CORPUS

Corpus

7 | 2008
Constitution et exploitation des corpus d'ancien et de moyen français

Constitution et exploitation des corpus d'ancien et de moyen français

Céline Guillot, Serge Heiden, Alexei Lavrentiev et Christiane Marchello-Nizia

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/1495
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 10 novembre 2008
ISSN : 1638-9808

Référence électronique
Céline Guillot, Serge Heiden, Alexei Lavrentiev et Christiane Marchello-Nizia, « Constitution et exploitation des corpus d'ancien et de moyen français », *Corpus* [En ligne], 7 | 2008, mis en ligne le 13 novembre 2009, consulté le 19 avril 2019. URL : http://journals.openedition.org/corpus/1495

Constitution et exploitation des corpus d'ancien et de moyen français

Céline Guillot, Serge Heiden, Alexei Lavrentiev et Christiane Marchello-Nizia

1. Introduction

1 Ce numéro spécial dédié aux corpus d'ancien et moyen français, à leur constitution et à leur exploitation, vient après bien d'autres numéros de revues présentantl'usage des corpus en linguistique depuis une dizaine d'années. Mais celui-ci est spécifiquement consacré aux corpus concernant les périodes les plus anciennes de la langue, c'est-à-dire aux formes et pratiques linguistiques qui sont les plus différentes de ce que connaît et pratique le locuteur moderne, et pour lesquelles la base textuelle de référence FRANTEXT ne fournit rien puisque les textes qu'elle comporte ne remontent pas au-delà de 1520.

2 Le but de ce numéro thématique est d'une part de dresser un état des lieux dans le domaine des corpus consacrés aux textes et documents de la période la plus ancienne du français (antérieure au XVI^{e} s.), mais aussi de développer une réflexion critique et épistémologique sur la nature et le rôle de ces corpus, dans trois de leurs aspects fondamentaux : sur leur structuration et leur mode de constitution d'une part, sur la façon dont ils ont pu infléchir aussi bien la méthodologie de la description des états anciens du français que la formulation d'hypothèses théoriques sur la langue plus généralement, et sur l'influence qu'ils ont pu avoir sur le développement et les avancées de nouveaux corpus plus récents.

3 Une spécificité des corpus de français du Moyen Age est leur nombre, relativement élevé, et leur dispersion tant dans l'espace que dans les objets qu'ils se sont fixés. Il est des raisons à ce polycentrisme : jusqu'en 2005 environ, plusieurs des corpus étaient en quelque sorte « privés », et donc difficilement accessibles aux chercheurs extérieurs ou étrangers au laboratoire qui les avait créés. D'autre part, le droit protégeant les intérêts

des éditeurs commerciaux français étant particulièrement strict, l'ouverture de ces corpus n'était guère encouragée. Enfin, plusieurs corpus se sont centrés sur une période (Base textuelle du moyen français) ou sur des textes reflétant la langue d'une certaine aire géographique (Anglo-norman Hub) ou sur un auteur et/ou la tradition manuscrite d'un texte (projet Charrette, projet Christine de Pizan). Le chercheur qui voudrait utiliser en 2008 un corpus d'ancien ou de moyen français a donc à sa disposition, selon son orientation ou ses besoins, plusieurs corpus bien différenciés. L'envers de cette richesse et de cette possibilité de choix est qu'il n'existe pas actuellement de corpus de référence pour ces périodes.

2. Une typologie des corpus de français médiéval existants

4 Nous donnerons d'emblée une liste et une description rapide des corpus de français médiéval actuellement disponibles. Certains ont été développés dès les années 80 du siècle dernier : avec le Corpus d'Amsterdam, Antonij Dees[1] fait figure de véritable pionnier puisqu'il a développé dès cette période un gros corpus diversifié et annoté. D'autres corpus sont beaucoup plus récents et en cours de développement : ainsicelui du Projet MCVF à Ottawa (Modéliser le changement : les Voies du français, resp. F. Martineau). Et au final, on constate une remarquable longévité de ces corpus, qui ont survécu et survivent aux aléas des politiques de recherche nationales et à la disparition de leur promoteur (c'est le cas du Corpus d'Amsterdam par exemple, qui est toutefois resté inaccessible pendant près d'une vingtaine d'années avant d'être rendu à la communauté des chercheurs grâce au travail d'A. Stein, P. Kunstmann, M. Gleßgen et de leurs équipes), et qui, tout en conservant leur spécificité originelle, travaillent de plus en plus en réseau, en particulier grâce à la création récente d'une structure couvrante, le CCFM (cf. ci-dessous). La présentation succincte que nous donnons de ces corpus dont le développement couvre près de trois décennies (pour une liste plus complète, voir Guillot, Lavrentiev, Marchello-Nizia 2007[2]) est organisée de façon à mettre en évidence les spécificités de chacun de ces ensembles. Il s'agit au total de sept corpus (neuf bientôt) : six rassemblent un grand nombre de textes sélectionnés suivant des critères différenciés, un l'édition d'un seul texte, mais à partir des divers manuscrits de ce texte, et deux sont en cours de développement.

5 Des six grands corpus de textes variés actuellement existants, un seul permet, outre l'interrogation en ligne, le téléchargement des textes complets librement et gratuitement :

a) Anglo-Norman On-line Hub (ANH)[3]

6 Ce site diffuse une version en ligne de l'*Anglo-Norman Dictionary* (AND) ainsi qu'une partie des textes sources cités dans le dictionnaire, en particulier les textes intégraux publiés par l'Anglo-Norman Text Society (ANTS). Ce projet est coordonné par David Trotter (Université d'Aberystwyth, Pays de Galles). Le corpus textuel comporte actuellement 75 textes intégraux encodés XML-TEI. Ces documents peuvent être téléchargés ou simplement consultés et interrogés en ligne grâce à un moteur de recherche (concordances KWIC).

7 Deux autres corpus, de grande ampleur, sont seulement interrogeables en ligne :

b) Base textuelle du Moyen Français (BTMF)[4]

8 Géré par l'équipe du moyen français du laboratoire ATILF de Nancy (resp. S. Bazin-Tacchella), ce corpus réunit un ensemble important de textes intégraux de moyen français (XIV^e^-XV^e^ siècles, 6 800 000 occurrences). Conçu et élaboré en complément du Dictionnaire du moyen français (DMF), il est librement interrogeable en ligne, dans les mêmes conditions et avec le même moteur de recherche (Stella) que la base FRANTEXT. Une seconde interface, spécifique à ce corpus, offre des possibilités de recherche dans les textes soit par forme soit par lemme grâce à un outil de lemmatisation associé à la base de textes.

c) Textes de Français Ancien (TFA)[5]

9 Constitué par le laboratoire de français ancien de l'Université d'Ottawa à l'initiative de Pierre Kunstmann, ce corpus rassemble des textes intégraux d'ancien et de moyen français (XII^e^-XV^e^ siècles), ainsi qu'un certain nombre de transcriptions de manuscrits (3 500 000 occurrences). Les textes sont hébergés à l'ARTFL, Université de Chicago, et interrogeables en ligne avec l'outil Philologic.

10 Un troisième type de corpus permet d'interroger en ligne des textes annotés.

d) Base de Français Médiéval (BFM)[6]

11 Créée en 1989 à l'initiative de C. Marchello-Nizia, cette base est actuellement hébergée à l'Ecole normale supérieure Lettres et Sciences humaines de Lyon (resp. C. Guillot). Elle donne accès à un corpus de 80 textes intégraux français, littéraires et non littéraires, composés entre le IX^e^ et le début du XVI^e^ siècle (plus de 3 000 000 occurrences). La période ancienne (avant 1300) est la mieux représentée et sera amenée à se développer encore[7]. Les textes peuvent être interrogés en ligne grâce à la plateforme textométrique Weblex développée par S. Heiden, par tout chercheur, enseignant ou étudiant qui en fait la demande (inscription gratuite) et accepte les conditions de son utilisation formulées dans la « Charte de la BFM »[8]. La base a par ailleurs été intégralement étiquetée en morphosyntaxe à l'aide du TreeTagger (étiquetage automatique non vérifié) et cinq textes ont fait l'objet d'un étiquetage semi-automatique (Sato) entièrement vérifié par les experts linguistes.

e) Corpus du projet Modéliser le changement : les voies du français (2005-2010, GTRC/Conseil de recherches en sciences humaines du Canada)[9]

12 Ce corpus réunit un ensemble de textes français, le plus souvent intégraux, écrits entre le XI^e^ et le XVIII^e^ siècle. Les textes sont organisés selon des critères dialectaux, sociaux et historiques (par grandes périodes). Ce corpus présente par ailleurs la particularité que les textes ont été ou seront intégralement étiquetés en morphosyntaxe, grâce à un jeu de 59

étiquettes, et annotés en syntaxe en association avec l'Université de Pennsylvanie. Pour l'instant, ce corpus n'est accessible qu'aux chercheurs associés au projet.

f) Nouveau Corpus d'Amsterdam (NCA)[10]

13 Il s'agit d'une version révisée (relue, lemmatisée et convertie en XML) du corpus constitué au début des années 1980 sous la direction d'Anthonij Dees (version électronique fournie par Piet van Reenen de l'Université libre d'Amsterdam) et qui a servi de base à l'*Atlas des formes linguistiques des textes littéraires de l'ancien français* (Dees1987). Ce corpus, qui mêle 300 extraits de 200 textes déjà édités et des transcriptions de manuscrits inédits (XII^{e}-XIV ^{e} siècles, plus de 3 000 000 d'occurrences), avait en outre été étiqueté manuellement en morphosyntaxe avec un jeu de 225 étiquettes. Le nouveau corpus d'Amsterdam, lemmatisé et diffusé par A. Stein (Université de Stuttgart), est téléchargeable (textes et outil d'exploitation) sur la toile après une inscription gratuite, une seconde version du NCA est depuis quelques mois accessible pour l'interrogation en ligne[11](accès restreint).

14 Enfin, un quatrième type de corpus offre des éditions en ligne de textes particuliers :

g) Projet Charrette 2[12]

15 Il s'agit d'une édition savante de la tradition manuscrite du *Chevalier de la Charrette* de Chrétien de Troyes. Les transcriptions des huit manuscrits existants ont été éditées en parallèle et l'édition critique du texte, élaborée par A. Foulet et K. Uitti[13], est enrichie d'un apparat critique important.

16 Le projet a été initié et dirigé jusqu'en 2003 par K. Uitti à l'Université de Princeton. Depuis 2004, le projet, baptisé désormais *Charrette 2*, n'est plus rattaché à une institution et est co-dirigé par R. Alvarado (Dickenson College), G. Greco (Université de Portland) et S.-J. Murray (Université Baylor).

17 L'outil d'exploitation en ligne Figura (développé par R. Alvarado) permet d'explorer les figures poétiques annotées dans le texte de l'édition critique (chiasmes, rimes riches, enjambement, etc.), d'effectuer des requêtes sur les mots du texte, entièrement étiquetés et lemmatisés et de parcourir côte à côte des pages des manuscrits sous forme de transcriptions et d'images en couleurs, ainsi quele texte de l'édition critique et la traduction en français moderne. Le texte de l'édition avec les annotations est téléchargeable, épisode par épisode, au format XML.

18 L'accès au corpus Charrette nécessite une inscription gratuite.

h) et i) Projets Graal (resp. C. Marchello-Nizia)[14] et Christine de Pizan (resp. James Laidlaw)[15]

19 Ces deux projets en cours de développement visent à mettre en ligne d'une part une édition multi-facettes du manuscrit de la *Queste del saint Graal* de la Bibliothèque municipale des Lyon (Palais des Arts 77), et d'autre part une édition du manuscrit le plus important de l'ensemble de l'œuvre de Christine de Pizan (British Library Harley MS 4431).

20 Ces corpus et bases textuelles ont été élaborés par des équipes appartenant à différentes institutions situées dans plusieurs pays. A de multiples reprises, les échanges ponctuels d'expériences et de textes ont permis l'intégration d'un même document dans plusieurs bases. Mais il n'est pas rare aussi qu'un même texte ait été numérisé simultanément par plusieurs équipes. Conscients d'une certaine dispersion des efforts et des ressources, les représentants de la plupart de ces projets (à savoir l'Université d'Ottawa, l'École normale supérieure Lettres et Sciences humaines, l'Université de Stuttgart, l'Université de Zürich, le laboratoire ATILF du CNRS et de l'Université Nancy 2, l'Université du Pays de Galles à Aberystwyth et l'École nationale des Chartes) ont mis en place en 2004 une structure fédérative, appelée *Consortium international pour les corpus de français médiéval* (CCFM, http://ccfm.ens-lsh.fr/) chargée de proposer des normes communes de description et d'encodage des textes, afin d'assurer une meilleure interopérabilité de ces différents corpus et, si possible, de donner à l'avenir aux chercheurs un point d'accès unique à l'ensemble des données disponibles.

3. Comment constituer les corpus médiévaux numériques aujourd'hui ?

21 La première question est : quoi numériser, pour qui, et à quelles fins ? Un corpus est toujours représentatif d'un usage de la langue. Lequel, lesquels sélectionne-t-on au départ ?

22 On sait que du Moyen Age tout ne nous est pas parvenu[16] : nombreux sont les textes que nous ne possédons que sous une forme incomplète ou « mutilée », l'exemple le plus célèbre étant sans doute celui des romans de Tristan, dont on ne possède plus que des fragments dispersés ; et parmi les six œuvres évoquées par Chrétien de Troyes dans le prologue de son roman *Cligés*, deux seulement nous sont connus : on aurait ainsi perdu cinq ouvrages de cet auteur, quatre adaptations d'œuvres ovidiennes et un roman *Du roi Marc et d'Yseut la Blonde*. Mais même pour les textes qui nous sont parvenus dans leur entier, combien de manuscrits ont-ils été perdus, ou brûlés dans des incendies ? Combien, justement parmi les plus anciens, ont-ils vu leur support – le parchemin en particulier – gratté pour être réutilisé et servir à une nouvelle copie, ou simplement découpé pour contribuer à rembourrer les couvertures de manuscrits plus récents ? Plus le temps passe, plus ces aléas deviennent rares. Mais pour les textes et documents les plus anciens en français, le choix est obligatoirement restreint. Et même si des XII^{e} et XIII^{e} s. un grand nombre de textes nous sont parvenus, c'est leur potentielle représentativité qui se trouve *de facto* réduite : la situation de « co-linguisme » (R. Balibar) ou de plurilinguisme réserve pendant des siècles au latin l'écriture des textes autres que de fiction. Pas de chartes ni de textes juridiques en français avant le XIII^{e} s., pas de texte philosophique avant la fin du XIII^{e} s., peu de textes historiques jusqu'au XIV^{e} s. L'impact de ces facteurs, propres aux conditions de production, de diffusion et de conservation des textes médiévaux, se réduit au fur et à mesure que l'on progresse dans le temps : augmentation très significative du nombre de manuscrits conservés à partir du XIII^{e} s. (cette augmentation étant directement liée à celle du nombre de manuscrits produits à partir de cette période), diversification croissante des types de textes produits (liée au développement de l'écrit et à la place qu'occupe le français, dans ses différentes variantes, face aux autres langues parlées et écrites).

23 Mais tous les textes qui nous ont été transmis malgré ces aléas et qui ont bénéficié d'une édition imprimée n'ont pas été systématiquement numérisés : des choix ont été opérés dans chacun des corpus évoqués ci-dessus, afin, à chaque fois, de correspondre aux visées définies pour chaque projet. Or la composition d'un corpus a bien évidemment un impact sur les résultats obtenus lors de requêtes effectuées sur ce corpus. Pour que ces résultats soient correctement interprétés, tous les éléments du choix initial doivent nécessairement être explicités et accessibles, en particulier sous la forme de « descripteurs » ou de métadonnées associées aux textes qui composent le corpus. De ce point de vue, la situation du français ancien n'est pas tellement différente de celle de tout autre état de langue : pour être exploitable, le corpus doit nous renseigner sur la typologie et la source des données qu'il met à disposition – mais elle est rendue plus problématique encore par la rareté de ces données.

24 Viendra un moment où l'évaluation de la représentativité des données numériques dont on dispose actuellement pour ces états anciens pourra et devra être faite : dans quelle mesure les résultats obtenus à partir de ces données sont-ils fiables, et jusqu'où permettent-ils d'aller dans la généralisation ?

25 Si la question du choix des œuvres est, comme on l'a vu, d'emblée biaisée pour des raisons historiques pour les siècles les plus anciens, il existe par ailleurs des critères qui président à une sélection raisonnée des textes à numériser pour les inclure dans un corpus. L'un des critères pris en compte presque systématiquement est la « qualité philologique » des textes « papier » : qualité des éditions choisies (fondées sur un manuscrit lui-même bien choisi, que l'éditeur suit le plus fidèlement possible sans trop le « corriger » ou l' « amender »), mais aussi finesse de leur description en termes de datation à la fois des œuvres et des manuscrits, de typologie des textes édités (domaine, genre, etc.), de coloration dialectale des textes provenant aussi bien de l'auteur que du copiste du manuscrit, etc.

4. Formatage, balisage et étiquetage : quel format choisir pour les corpus numériques ?

26 On entend souvent parler du « texte brut », opposé au texte « avec la mise en forme » (dans le domaine de la bureautique) ou « balisé » (dans le domaine de la linguistique de corpus).

27 La notion du texte brut n'est pas aussi triviale qu'elle peut le paraître : un texte français privé de toute mise en forme ne serait qu'une chaîne de lettres difficilement lisible (à l'image de la *scripta continua* de manuscrits antiques grecs et latins). Les signes de ponctuation, les majuscules et les espaces blancs introduits par l'éditeur moderne sont en effet déjà des marques de structuration ou des « balises implicites » en quelque sorte ; et même si elles apparaissent dans les manuscrits, leur usage et leur signification ne sont pas les nôtres et doivent faire l'objet d'une « édition » au même titre que le reste.

28 Bref, même si on considère habituellement les espaces blancs, les sauts de ligne et les marques de ponctuation – de l'éditeur moderne aussi bien que du copiste médiéval – comme faisant partie du texte brut, les outils informatiques du traitement automatique de la langue (TAL) considèrent ces caractères comme des séparateurs plus ou moins forts permettant de construire une première couche de modélisation de la structure du texte.

29 Mais la réalité du texte écrit ou imprimé est bien plus complexe : on peut distinguer au moins deux structures parallèles, la structure sémantique (graphèmes – mots – phrases – divisions du texte) et la structure physique (ligne – page – volume), sans compter les différentes mises en relief, etc.

30 Pour l'ordinateur, il est nécessaire de coder explicitement toutes ces structures, si l'on veut pouvoir en tenir compte dans l'exploitation d'un corpus. Une solution technologique née dans les années 1960 et ayant en 1996 abouti à la forme d'un « langage de balisage » (*markup language*) XML, consiste à séparer le « texte brut » (chaîne de caractères et d'espaces blancs) des « balises » portant toutes sortes de métadonnées. Ce sont donc ces balises qui permettent d'encoder la structure (ou plutôt les différentes structures) qui organisent le texte, mais aussi de doter le texte d'un apparat critique, d'annotations linguistiques et philologiques, de marques typographiques, de descripteurs bibliographiques, etc.

31 La norme XML ne pose en effet que quelques contraintes relatives à la forme et à la « syntaxe formelle » du balisage, mais laisse à l'encodeur la liberté de sélectionner les métadonnées (dont le nom, la place et l'interprétation) qu'il souhaite associer à son texte brut.

32 La TEI (*Text Encoding Initiative*) vise à son tour à spécifier plus précisément le jeu de balises et leurs règles d'utilisation dans tous types d'éditions électroniques de textes afin de faciliter l'échange des données et la mutualisation des outils qui les manipulent. Même si les recommandations de la TEI, qui reposent sur une vingtaine d'années d'expérience dans le domaine du codage des textes, offrent un guide précieux pour les principes du balisage, chaque nouveau projet d'édition électronique ou d'un corpus textuel est amené à faire des choix entre les nombreuses possibilités qui existent.

33 Si certains éléments semblent incontournables (par exemple baliser les sauts de page en cas de numérisation d'une édition imprimée, ou encore les paragraphes dans des textes en prose), d'autres choix de balisage sont moins évidents. Faut-il privilégier la structure formelle (la division en vers, par exemple) ou la structure linguistique (la division en phrases)[17] ? Comment relier les annotations au corps du texte (cf. la problématique de l'annotation « débarquée » dans l'article de N. Mazziotta) ? Quel jeu d'étiquettes choisir pour un enrichissement linguistique (cf. la section sur les corpus enrichis dans l'article de S. Prévost et les principes d'annotation syntaxiques dans l'article de A. Stein) ? Quels descripteurs associer au texte pour faciliter son intégration dans des corpus et son référencement dans des bases de données bibliographiques ?

34 Les questions, importantes, du choix des balises XML et de l'adaptation, nécessaire dans chaque cas, des recommandations de la TEI, sont abordées dans plusieurs articles de ce numéro (Martineau, Mazziotta, Stein) et sont au cœur des discussions menées au sein du CCFM.

35 Apparaîtra aussi, par exemple, la constatation que les difficultés rencontrées dans la phase d'étiquetage des items du texte ont pour contrainte la mise au point préalable d'un jeu de catégories cohérent et couvrant, ce qui signifie une clarification des notions et concepts utilisés dans la définition et la catégorisation des items (mots, groupes de mots, phrases, mais aussi, s'agissant de textes qui nous ont été transmis uniquement par des manuscrits copiés par des scribes bénéficiant d'une certaine latitude de liberté et d'interprétation, graphies, ponctuation, paragraphage). Une fois le jeu défini, on peut alors éprouver des difficultés à représenter le niveau d'enrichissement : par exemple Dees

en 1980, Ollier en 1990 avaient choisi de coder les étiquettes en chiffres ; chez Dees l'étiquette « 155 » s'interprète comme « article démonstratif sujet féminin singulier ».

36 Mais ces difficultés ne doivent pas occulter la richesse de ce type de traitements qui ouvrent des possibilités de recherche tout à fait nouvelles et novatrices et qui nous renseignent, par les éléments qui sont ainsi mis en évidence, sur ces états de langue passés.

5. Quels sont les impacts de l'usage des corpus numériques ?

37 De fait, l'existence et la possibilité d'exploiter plus ou moins automatiquement des corpus ont eu pour conséquence un changement fondamental dans l'analyse linguistique des états de langue en question – sans locuteur natif[18], mais aussi dans les études historiques, sémiotiques et littéraires.

38 En linguistique en particulier, on ne peut plus faire l'économie d'une approche « corpussielle » ; de ce point de vue nous sommes entrés, épistémologiquement, dans un nouvel âge[19]. En effet, ce nouveau type d'approche permet – contraint à – deux avancées capitales dans le champ des analyses linguistiques.

39 D'une part les « observables » linguistiques sont objectivés par 1) une explicitation au sein du corpus de ce qui est pris en compte dans l'analyse (ce qui est cherché et listé, ce qui est compté – y compris le niveau auquel se situe l'analyse linguistique), 2) le paramétrage, au moyen de descripteurs explicites, du périmètre sociolinguistique de ce qui est pris en compte dans l'analyse (ce qui permet de relativiser, comparer, contraster simultanément dans le temps, les aires géographiques, etc.), 3) l'échange des observables grâce à la compatibilité désormaisrendue possible, et le cumul des acquis par le biais de l'échange ou de l'accès aux mêmes corpus. La rigueur et la transparence initiales nécessaires dans l'établissement des données et dans leur description assurent une plus grande fiabilité et par la suite une comparabilité des résultats, et facilitent un effet cumulatif non seulement des acquis, mais aussi des corpus eux-mêmes.

40 D'autre part – et c'est là un point essentiel dans les études diachroniques – la quantification des faits langagiers devient opératoire : désormais, la récurrence d'apparition, ou d'absence, peut prendre une valeur heuristique voire probatoire pour certaines études linguistiques.

6. Présentation des articles

41 Ce numéro de CORPUS comprend huit contributions, dont chacune développe un aspect important de la problématique que nous venons d'évoquer[20].

42 L'article de Sophie Prévost, intitulé *Corpus informatisés de français médiéval, contraintes sur leur constitution et spécificités de leurs apports*, aborde de façon synthétique les problèmes centraux que posent la constitution et l'exploitation des corpus médiévaux. Il replace ces questions dans le cadre général du développement des recherches en histoire de la langue et fait état des raisons qui expliquent que ces recherches aient une longue tradition d'étude « sur corpus ».

43 Les deux articles suivants, *Lexicographie du français médiéval et corpus informatisé, le traitement lexicographique de verbes d'opinion et de connaissance à partir du corpus du DMF* de Corinne Féron et Parce que, perché, porque *dans les langues romanes médiévales : l'utilité des études sur corpus* de Benjamin Fagard présentent et illustrent les apports et limites des corpus existants, dans les domaines qui peuvent paraître les plus simples : le lexique et la morphologie, dont les items sont facilement identifiables par des chaînes de caractères. En suivant les différentes étapes du travail de description lexicographique, Corinne Féron montre en quoi les corpus numériques ont profondément modifié certaines pratiques lexicographiques, tout en pointant les limites des outils actuels et les perspectives d'évolution qui lui semblent souhaitables. Benjamin Fagard adopte quant à lui une perspective comparative, en abordant d'une part les spécificités et la complémentarité des outils traditionnels et des corpus numériques, et en présentant d'autre part une recherche sur l'expression de la cause dans trois langues romanes (français, italien, espagnol), menée grâce à l'exploration de trois bases de données différentes. Son article donne par ailleurs en annexe un tableau synthétique des principales bases de données disponibles pour les langues romanes.

44 Les deux contributions suivantes s'intéressent à la question de la variation linguistique et à sa mise en relation avec les conditions de production et de réception des documents médiévaux. Les bases de données enrichies de descripteurs, qui peuvent donner des informations sur l'origine sociale et ou géographique de l'auteur, du (des) copiste(s) du manuscrit, mais aussi sur les types de textes représentés dans la base, etc., permettent l'élaboration de sous-corpus ciblés et donc l'analyse comparative entre dialectes, entre langue standard et dialecte, ou entre différents types de textes, etc. L'étude de Richard Ingham, intitulée *The Grammar of later medieval French: an initial Exploration of the Anglo Norman Dictionary Textbase,* et celle de France Martineau, intitulée *Un corpus pour l'analyse de la variation et du changement linguistique*, illustrent ces possibilités, en insistant sur les limites des recherches qui sont actuellement possibles (R. Ingham) et sur les conditions nécessaires au développement de ces exploitations (F. Martineau). Ces différents points sont illustrés à partir de l'étude de phénomènes linguistiques précis : négation, inversion du sujet, article partitif, etc. Dans sonarticle, France Martineau aborde également la problématique de l'enrichissement linguistique des données numériques, en particulier dans le cadre de l'annotation syntaxique pratiquée dans le projet *Modéliser le changement : les Voies du français* (Université d'Ottawa), question qui est au centre de l'étude suivante.

45 L'article d'Achim Stein, intitulé *Syntactic Annotation of Old French Text Corpora*, relate en effet l'expérience réalisée sur le Corpus d'Amsterdam, créé il y a vingt ans par Antonij Dees, revivifié récemment (Nouveau Corpus d'Amsterdam), et sur lequel on projette de créer une nouvelle couche d'annotation linguistique (syntaxique). Le problème de la procédure et du format d'annotation, qui est abordé icide façon relativement détaillée dans une perspective comparative avec les projets existants ou en cours pour le français et l'anglais médiéval, montre notamment l'importance de la théorie linguistique sous-jacente à l'outil d'annotation : ainsi par exemple, selon la théorie adoptée, on annotera, ou non, une place vide quand le sujet n'est pas exprimé – ce qui induit des stratégies de requête différentes pour les utilisateurs.

46 L'article qui suit, La quête du Graal *et la réalité numérique*, présente l'expérience menée par une équipe d'informaticiens-linguistes-littéraires (Claire Serp, Anne Laurent, Mathieu Roche, Maguelonne Teisseire) sur le très gros corpus homogène du Lancelot-Graal. Cette recherche, qui vise à utiliser différents outils de TAL dans le cadre d'une étude sur les

noms de parenté, illustre une démarche qui se veut au départ non ciblée, afin de conserver à l'empirie toute sa valeur heuristique – c'est ainsi que s'est révélée, de façon inattendue, l'importance conceptuelle d'un nom *a priori* 'vide', le mot *chose.*

47 Le dernier article de ce numéro, *Traiter les abréviations du français médiéval. Théorie de l'écriture et pratiques d'encodage* de Nicolas Mazziotta, montre combien la frontière entre le « technique » et le « scientifique » est ténue. Une recherche, dont le point de départ est un problème d'ordre pratique et formel (le codage des abréviations médiévales), peut ainsi déboucher sur la définition d'un classement typologique et sur la description d'un système complexe.

7. Conclusion : un nouveau dialogue entre la philologie et les éditions numériques ?

48 De façon générale, la pratique linguistique des médiévistes a été notablement modifiée, comme le montrent les différents articles de ce numéro, par la possibilité (et même désormais l'obligation) de recourir à ces outils, et à ces nouvelles formes de représentation des textes.

49 Mais l'on voudrait souligner, en conclusion, un problème important désormais pour les médiévistes, non spécifiquement lié à la constitution des corpus, mais qui en est une conséquence : c'est la question du transfert des éditions savantes dans le monde électronique – c'est-à-dire, à terme, la transformation des normes éditoriales sous l'influence des apports des nouveaux outils, comme par exemple dans l'édition GRAAL multi-couches (voir Lavrentiev 2005[21] et Haugen 2004[22]), avec la possibilité offerte à chaque utilisateur de choisir son niveau de représentation, qui peut être différent selon le thème ou l'objet qu'il traite.

50 Jusqu'à tout récemment, avec les éditions papier ou la simple lecture des manuscrits, une seule version, et donc un point de vue unique sur le texte s'imposait. Désormais la forme sous laquelle nous apparaît un document n'est plus nécessairement « donnée », mais elle peut – et parfois doit – être choisie, « construite » par le lecteur ou l'utilisateur lui-même. C'est ce qui se passe pour les éditions en ligne du type de celles qu'offrent le Projet Charrette 2 ou le Projet Graal (ENS-LSH Lyon).

51 Le choix se situe désormais à deux niveaux au moins : au plan du *mode de visualisation* de ces données (on peut donner plusieurs vues du même document en fonction des utilisateurs et de besoins différents) et cela à tous les niveaux (texte, parties de texte, mots, caractères, etc.) ; et au plan de la *sélection des informations* que l'on souhaite voir apparaître.

52 Si de telles possibilités techniques ont pu être développées et peuvent être exploitées, c'est que, dès ses origines avec l'étude des textes anciens, la philologie « traditionnelle » a toujours été consciente du travail d'interprétation que comporte l'établissement du texte. Et cela a d'emblée rendu possible un dialogue avec la linguistique de corpus, comme l'illustre par exemple l'introduction de Jacques Monfrin à l'ouvrage de Claire Blanche-Benveniste et Colette Jeanjean (1987), *Le français parlé. Transcription et édition*, Paris, Didier Erudition.

53 Tout choix, dans la perspective de ces éditions dynamiques, implique une intervention, qu'il s'agisse de celle du concepteur ou de celle de l'utilisateur – et c'est ce dernier aspect

qui est nouveau –, donc une information et une réflexion préalables, des compétences, et une explicitation des buts que l'on poursuit.

54 Les éditions électroniques auront en particulier l'avantage de pouvoir s'intégrer sans un effort trop important et coûteux dans une base de textes ou un corpus[23], à condition d'être balisées et formatées de façon normée, ce qui pose le choix des standards de formatage, l'homogénéisation des pratiques et leur institutionnalisation, car telles sont les conditions de leur interopérabilité. Et à cet égard, le CCFM joue pleinement depuis deux ans son rôle d'échange de pratiques et de normes de description et de codage entre partenaires médiévistes, ainsi que de diffusion en son sein des normes plus généralistes, comme celle du consortium TEI.

55 La pratique croissante des éditions électroniques obligera certainement, par ailleurs, à reformuler tout un pan du droit des auteurs et des éditeurs commerciaux vis-à-vis de la propriété intellectuelle véhiculée au sein des corpus à différents niveaux de granularité et pour différents niveaux d'analyse – mais là n'est pas notre propos.

56 C'est au prix de ces transformations fondamentales et de différente nature que l'on pourra bénéficier facilement des atouts de la lecture hypertextuelle instrumentée, et d'une circulation optimisée des fruits du travail des éditeurs de textes. Éditeurs, faut-il le rappeler, qui sont les premiers utilisateurs des corpus ainsi constitués.

57 Le renouvellement des questionnements sur les pratiques d'édition des manuscrits[24] aura tout à gagner à se faire à travers un dialogue avec la tradition des éditions philologiques, et sans doute dans leur lignée, car leurs procédures et leurs méthodes, nourries de plus de deux siècles d'une pratique très riche et d'une réflexion constante (on connaît au moins les débats suscités par les prises de position de K. Lachmann au début du XIX^{e} s., ou de J. Bédier un siècle plus tard), rencontrent tout un pan des questions soulevées par l'édition électronique.

NOTES

1. Dees A. (éd.) (1980) Atlas des formes et des constructions des chartes françaises du 13ème siècle, Tübingen : Niemeyer et Dees A. (éd.) (1987) Atlas des formes linguistiques des textes littéraires de l'ancien français. Tübingen : Niemeyer.

2. Guillot C., Lavrentiev A. & Marchello-Nizia C. (2007). « Les corpus de français médiéval : état des lieux et perspectives », in : Revue fançaise de linguistique appliquée 121 : 125-128. Il ne nous a pas paru souhaitable, dans le cadre d'un numéro consacré aux corpus de français ancien, de donner une liste exhaustive de tous les matériaux actuellement disponibles pour cette période. Nous nous limiterons dans cette présentation aux projets d'une ampleur et d'une richesse suffisante.

3. http://www.anglo-norman.net/

4. http://atilf.atilf.fr/dmf.htm

5. http://www.lib.uchicago.edu/efts/ARTFL/projects/TLA/

6. http://bfm.ens-lsh.fr

7. Le projet CORPTEF (« Corpus représentatif des premiers textes français »), qui a récemment obtenu un financement de l'Agence nationale pour la recherche vise en effet à constituer et exploiter un corpus aussi diversifié que possible de textes antérieurs au XIII^{e} siècle.

8. La base est toutefois momentanément fermée en raison d'un différend avec une maison d'édition. Elle devrait rouvrir prochainement, moyennant quelques modifications techniques de l'outil et la limitation de certaines possibilités d'interrogation.

9. http://www.voies.uottawa.ca/

10. http://www.uni-stuttgart.de/lingrom/stein/corpus/#nca

11. http://corpora.philosophie.uni-stuttgart.de/nca

12. http://lancelot.baylor.edu/

13. Paris, Bordas, 1989. Le texte de l'édition originale a été revu et corrigé par Karl Uitti dans le cadre du projet Charrette.

14. http://weblex.ens-lsh.fr/pub/kq

15. http://www.pizan.lib.ed.ac.uk/

16. Nous n'insisterons naturellement pas sur le problème de l'absence de données orales pour cette période de l'histoire du français, les seuls témoins que nous possédons – les passages au discours direct dans les textes narratifs ou les textes dramatiques – n'étant que des « représentations d'oral » et donc à utiliser avec la plus grande prudence de ce point de vue.

17. Le choix d'une structure arborescente unique pour un document est imposé par la norme XML. Il existe de multiples solutions permettant d'encoder des structures concurrentes, mais le choix d'une structure principale est nécessaire.

18. Marchello-Nizia C. (1985) « Questions de méthode », Romania, 106 : 481-492.

19. Marchello-Nizia C. (2004) « Linguistique historique, linguistique outillée : les fruits d'une tradition ». In C. Fuchs & B. Habert (éds.), Traitements automatiques et ressources numérisées pour le français. Le français moderne, volume 72/1 : 58-70. Prévost S. (2005) « Exploitation d'un corpus de français médiéval : enjeux, spécificités et apports ». In A. Condamines (éd.), Sémantique et corpus.Paris : Hermès/Lavoisier, pp. 147-176.

20. Nous remercions les collègues qui ont accepté d'être les relecteurs de ces articles, et spécialement Bénédicte Pincemin, Marie-Hélène Lay, Fernande Dupuis, David Trotter, Bernard Combettes, Sylvie Mellet.

21. Lavrentiev A. (2005) « Représentation de transcriptions diplomatiques de manuscrits français médiévaux en XML-TEI ». In J. Kabatek, C. Pusch & W. Raible (éds.) Romanistische Korpuslingustik II: Korpora und diachrone Sprachwissenschaft, Romance Corpus Linguistics II: Corpora and Diachronic Linguistics. Tübingen : Gunter Narr Verlag, (ScriptOralia ; 130), pp. 109-121.

22. Haugen O. E. (2004) « Parallel Views: Multi-level Encoding of Medieval Nordic Primary Sources », Literary and Linguistic Computing 19, 1 : 73-91.

23. Une interopérabilité totale et automatique entre différents corpus semble à ce jour une gageure, mais le respect des normes internationales et la documentation des pratiques d'encodage et de description peuvent faciliter considérablement les projets d'échanges de données et d'intégration de corpus.

24. Questionnements, il faut le rappeler, qui furent initiés d'abord par P. Zumthor en 1972 avec la notion de la « mouvance » des textes (Essai de poétique médiévale, Paris, Le Seuil), puis en 1989 par l'ouvrage de B. Cerquiglini Eloge de la variante, Histoire critique de la philologie (Paris, Le Seuil) et en 1990 dans un numéro spécial de la revue Speculum composé par les médiévistes américains se réclamant de la New Philology (en particulier l'article de R. H. Bloch intitulé « New Philology and Old French »). Ces réflexions restaient largement programmatiques, mais certaines d'entre elles peuvent nourrir encore aujourd'hui notre réflexion et notre pratique.

AUTEURS

CÉLINE GUILLOT

ENS-LSH Lyon, UMR 5191 ICAR

SERGE HEIDEN

ENS-LSH Lyon, UMR 5191 ICAR

ALEXEI LAVRENTIEV

ENS-LSH Lyon, UMR 5191 ICAR

CHRISTIANE MARCHELLO-NIZIA

ENS-LSH Lyon, UMR 5191 ICAR