Développement de la Base de français médiéval : qualité philologique, ouverture et outillage textométrique

Alexei Lavrentiev

A. Lavrentiev (Lyon)

DÉVELOPPEMENT DE LA BASE DE FRANÇAIS MÉDIÉVAL : QUALITÉ PHILOLOGIQUE, OUVERTURE ET OUTILLAGE TEXTOMÉTRIQUE

This paper presents three key aspects in the development of the Base de Français Médiéval Old French text corpus: the quality of data, open source policy for texts and software, and improvement of tools for reading, searching and and analyzing the texts of the corpus.

В докладе представлены три ключевых аспекта развития Базы средневекового французского языка (BFM): повышение качества данных, открытость текстов и программного обеспечения и улучшение инструментов для чтения, поиска и анализа текстов корпуса.

La Base de français médiéval (BFM) a été créée à l'initiative de Christiane Marchello-Nizia en 1989. Son objectif était de permettre aux chercheurs et aux étudiants de profiter des technologies informatiques pour l'étude de la langue et de la littérature françaises médiévales. L'ordinateur permet notamment de retrouver rapidement toutes les occurrences d'une forme ou d'une construction dans un grand nombre de textes, ce qui accélère énormément le travail du chercheur. À l'époque, une grande base textuelle informatisée (Frantext) existait déjà pour le français moderne (à partir du XVI^{e} siècle) et le travail était bien avancé sur la constitution d'une base textuelle pour le *Dictionnaire du moyen français* (DMF). Les deux projets étaient développés à l'Institut national de la langue française à Nancy (actuellement UMR ATILF, http://www.atilf.fr). Il s'agissait donc de compléter ces ressources pour la période la plus ancienne de l'histoire du français (des premiers textes au début du XIV^{e} siècle). Certains choix méthodologiques (utilisation d'éditions scientifiques de référence et non de manuscrits, élimination de l'apparat critique, inclusion de textes intégraux) ont hérité de ces projets antérieurs.

Au cours de plus d'un quart de siècle de son histoire (un âge respectable dans le monde du numérique!), la BFM a su s'adapter aux évolutions technologiques, méthodologiques et sociétales pour offrir à la communauté des médiévistes une ressource toujours plus riche, plus fiable et plus ouverte. À ce jour, la BFM comporte 126 textes (soit 3 550 000 mots), datés du IX^{e} au XV^{e} siècle et constitue ainsi l'un des corpus numériques les plus importants pour la période médiévale de l'histoire du français. Accessible en ligne via un portail TXM (http://txm.bfm-corpus.org), ce corpus se compose d'éditions de référence (éditions originales et éditions imprimées numérisées), encodées au format XML-TEI et enrichies à de multiples niveaux: métadonnées décrivant les textes, codage interne aux textes et segmentation graphique (tokenisation), étiquetage morphosyntaxique, annotation syntaxique, encodage du discours direct. Près de 400 utilisateurs sont actuellement inscrits au portail de la BFM.

Trois aspects de développement de corpus textuels nous paraissent essentiels

pour que ces ressources puissent modifier profondément (sinon révolutionner) la recherche en sciences humaines et sociales. Il s'agit de la qualité des données, de leur ouverture et des outils d'analyse mis à disposition des chercheurs. Dans notre communication, nous aborderons chacun de ces aspects dans l'historique, dans l'état actuel et dans les perspectives de la BFM.

La qualité des données est une question cruciale pour toute recherche basée sur corpus numérique. Si les données ne sont pas fiables ou si leur qualité n'a pas été évaluée, il n'y a aucune certitude sur la valeur des résultats de recherche obtenus sur ces données. Selon Lydia Stanovaïa, «un chercheur visant à obtenir les données réelles doit absolument étudier les manuscrits réels, existants, conservés, faits à l'époque étudiée, et non pas les textes irréels, hypothétiques, reconstruits par des philologues» [Stavovaïa 2003: 246]. Les éditions «critiques» seraient donc inutilisables dans le cadre de recherches linguistiques. Le point de vue que nous adoptons est moins radical. Sans utiliser les éditions scientifiques modernes, il aurait été impossible de construire un corpus textuel important en un temps et au prix raisonnables. Si les éditions choisies sont de type «bédiériste» [Bédier 1928] (comme c'est le cas de la majorité d'éditions de textes en français médiéval) et si elles sont établies avec la rigueur philologique nécessaire, elle sont tout-à-fait utilisables pour les recherches littéraires, historiques et plus la plupart des recherches linguistique, à l'exception des travaux sur la segmentation graphique, la ponctuation et l'usage des abréviations [Lavrentiev 2007]. Bien entendu, la «fidélité» de chaque édition doit être étudiée de près avant son intégration dans un corpus de recherche linguistique. La «normalisation» opérée par les éditeurs scientifiques offre par ailleurs des avantages par rapport aux transcriptions «pures» des manuscrits: certaines formes graphiquement ambiguës dans les manuscrits sont distinguées (*parle* vs *parlé*, *jeut* vs *i eut*), ce qui facilite considérablement la lecture, rend la recherche de formes plus précise et améliore les performances de l'étiquetage moprhosyntaxique automatique. L'usage des guillemets par l'éditeur permet de repérer les passages au discours direct, même s'il faut toujours tenir compte des erreurs d'interprétation et des ambiguïtés possibles.

Dans le cadre d'éditions «nativement numériques», comme celle de *Queste del saint Graal* [La Queste 2013], le dilemme de normalisation *vs* fidélité à la source peut être résolu grâce aux transcriptions «multi-facettes» où l'utilisateur peut choisir le niveau de normalisation souhaitée. En plus, l'image du manuscrit peut être affichée à côté de la transcription pour permettre la vérification ce cette dernière. Depuis plusieurs années, l'équipe de la BFM travaille à la mise en place d'une collection d'éditions numériques originales (inspirée de l'expérience de la *Queste*) qui sera basée sur ces principes philologiques de transcriptions multi-facettes et de mise à disposition des images des sources. Plusieurs projets d'éditions de ce type sont en cours de réalisation et certaines d'entre elles sont déjà accessibles en tant que prototypes [Pignatelli/Lavrentiev sous presse; Lavrentiev/Markova 2014]. Progressivement, ces nouvelles éditions prendront une place de plus en plus importante dans le corpus de la BFM [Guillot *et al.* sous presse].

L'ouverture des données et des outils d'analyse de corpus est tout aussi importante que la qualité des données. Les éditions traditionnelles imprimées posent

des problèmes complexes liés à la propriété intellectuelle, qui empêchent le développement et la libre circulation de ressources numériques basées sur ces éditions [Guerreau 2015]. Or l'accès libre aux données de recherche, l'archivage ouvert, leur réutilisation et leur enrichissement collaboratifs sont les meilleures, sinon les seules garanties de pérennité dans un monde du numérique en constante évolution technologique. L'utilisation de formats ouverts et conformes aux standards internationaux est également indispensable pour limiter le risque de perte de données (et parfois des années de travail). La BFM utilise depuis le début des années 2000 le format XML et le balisage conforme aux recommandations du consortium *Text Encoding Initiative* (http://www.tei-c.org) et met à disposition de ses utilisateurs l'ensemble de ses textes sous une licence ouverte *Creative Commons* «Attribution – Pas d'Utilisation Commerciale – Partage dans les Mêmes Conditions» (http://creativecommons.org/licenses/by-nc-sa/3.0/fr).

La «révolution numérique» dans la linguistique diachronique serait impossible sans le développement d'outils d'analyse de corpus, qui sont à la fois puissants et maniable par les chercheurs en sciences humaines et sociales. À ses débuts, la BFM était diffusée sur CD-ROM, en forme de concordances imprimables au format Microsoft Word. En 2002, la BFM devient interrogeable en ligne via le logiciel lexicométrique Weblex (développé par Serge Heiden). Ce logiciel proposait de très nombreuses fonctionnalités d'analyse qualitative et quantitative, mais son usage était difficile pour les non spécialistes en linguistique de corpus. Certaines fonctionnalités très demandées (comme la création de sous-corpus «sur mesure») étaient absentes.

Depuis 2012, la BFM est diffusée via un portail web réalisé sur la plateforme TXM (logiciel textométrique de nouvelle génération qui succède à Weblex et plusieurs autres logiciels [Heiden *et al.* 2010], http://sourceforge.net/projects/txm). Pour interroger la BFM il suffit de s'inscrire gratuitement au portail (la procédure ne prend que quelques instants). Sans inscription, les internautes peuvent lire les textes de la BFM en ligne, les télécharger au format PDF et encore accéder à leurs notices bibliographiques (métadonnées). Les utilisateurs inscrits peuvent créer des sous-corpus et des partitions (pour l'analyse contrastive), afficher des index et des concordances de formes et de catégories grammaticales et les exporter pour annotation ou citation dans des publications.

Des fonctionnalités d'analyse statistique (calculs de cooccurrence, de spécificités, construction de plans factoriels sont implémentés dans la «version bureau» de TXM (installable «localement» sur un ordinateur Linux, Windows ou Mac). A cette fin, une version «binaire» de la BFM qu'on peut facilement charger dans TXM «bureau» est proposée sur demande aux utilisateurs.

La documentation et l'assistance aux utilisateurs prennent une part importante dans le développement de la BFM. Les principales fonctionnalités du portail BFM sont décrites dans un «Tutoriel» accessible en ligne [Bertrand *et al.* 2014]. Un tutoriel vidéo est proposé pour l'édition de la *Queste del saint Graal* qui fait partie du portail de la BFM. Une «Foire aux questions» et une liste de diffusion sont mises à disposition des utilisateurs pour répondre aux problèmes méthodologiques et techniques les plus courants. Des formations gratuites à la plateforme TXM sont régulièrement organisées à Lyon et dans le cadre d'écoles thématiques diverses.

La BFM continue de se développer et de s'enrichir. De nouvelles fonctionnalités sont régulièrement proposées et les fonctionnalités existantes sont améliorées afin de rendre le travail avec la Base plus simple et efficace. Les chercheurs et les étudiants en linguistique et philologie romanes sont non seulement invités à s'inscrire et à interroger la BFM dans le cadre de leur travaux de recherche, mais aussi à contribuer à son développement en signalant les erreurs éventuelles et en participant à l'annotation des textes (étiquetage morphosyntaxique, annotation syntaxique, annotation du discours direct, etc.) et à la préparation de nouvelles éditions numériques.

Références bibliographiques

Bédier J. La tradition manuscrite du Lai de l'Ombre, réflexions sur l'art d'éditer les anciens textes // Romania, vol. 54, 1928, p. 161-196; 236-356.

Bertrand L. et al. Tutoriel TXM pour la BFM. Version 2.0. Lyon: ENS de Lyon, 2014. [http://txm.bfm-corpus.org/files/Tutoriel_TXM_BFM_V1.pdf].

Guerreau. A. L'avenir de la philologie textes anciens et domaine public. Paris: HAL, 2015. [https://halshs.archives-ouvertes.fr/halshs-01112213].

Guillot C. et al. La «philologie numérique»: tentative de définition d'un nouvel objet éditorial // Actes du XXVII^{e} Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013) / Éd. Buchi É., Chauveau J.-P. et Pierrel J-M. 3 volumes. Strasbourg: Société de linguistique romane/ÉliPhi, sous presse. [https://halshs.archives-ouvertes.fr/halshs-00846767].

Heiden S. et al. TXM: Une plateforme logicielle open-source pour la textométrie – conception et développement. // 10th International Conference on the Statistical Analysis of Textual Data – JADT 2010 (June 2010, Rome, Italy) / Éd. Bolasco S., Chiari I. et Giuliano L. Vol. 2. Rome: Edizioni Universitarie di Lettere Economia Diritto, 2010. p.1021-1032. [https://halshs.archives-ouvertes.fr/halshs-00549779]

La Queste del saint Graal: Édition numérique interactive du manuscrit Lyon, BM, P.A. 77 / Éd. Marchello-Nizia C., Lavrentiev A. Lyon: ENS de Lyon, 2013. [http://catalog.bfm-corpus.org/qgraal_cm].

Lavrentiev A. Base de français médiéval et transcriptions de manuscrits: recherche de complémentarité // Actes du XXIV^{e} Congrès International de Linguistique et de Philologie Romanes (Aberystwyth, 1-6 aout 2004) / Éd. Trotter D. Tübingen: Max Niemeyer, 2007. p. 405-410.

Lavrentiev A., Markova E. From the Holy Grail to the Good Health: a Digital Edition of a 15th Century French Medical Treatise on the BFM Web Portal // Textual Heritage and Information Technologies. El'Manuscript - 2014, Sep 2014 / Éd. Baranov V. et al. Sofia, Izhevsk: Cyrillo-Methodian Research Center and Izhevsk University, p. 164-166. [https://halshs.archives-ouvertes.fr/halshs-01071459].

Pignatelli C. et Lavrentiev A. Le Psautier d'Arundel : une nouvelle édition // Actes du XXVII^{e} Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013) / Éd. Buchi É., Chauveau J.-P. et Pierrel J-M. Nancy: ATILF, sous presse. [https://halshs.archives-ouvertes.fr/halshs-00846770].

Stanovaïa L. A. La standardisation en ancien français // The Dawn of the Written Vernacular in Western Europe / Éd. Goyens M. et Verbeke W. Leuven: Leuven University Press, 2003. p. 241-272.