# Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique

Céline Guillot, Serge Heiden, Alexei Lavrentiev

# Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique[1]

Céline Guillot-Barbance
Serge Heiden
Alexei Lavrentiev
UMR IHRIM – ENS de Lyon / Université de Lyon / CNRS – Labex Aslan

## Introduction

Les mutations induites par le développement du numérique dans le champ des Sciences du langage ont eu une répercussion très directe ces dernières années sur la linguistique diachronique, tout spécialement dans le domaine français. Par son objet d'étude – les états de langue passés pour lesquels nous ne disposons pas de locuteurs ni de compétence linguistique –, la linguistique diachronique s'appuie depuis toujours sur des corpus de données attestées. Mais l'essor récent des ressources numériques a considérablement renouvelé les méthodologies d'analyse, les résultats produits par la recherche et parfois aussi les phénomènes étudiés. Ces évolutions en cours tendent à renforcer l'attitude réflexive du diachronicien, par nécessité confronté à l'altérité des données qu'il décrit. Et par certains côtés, les questions nouvellement posées par l'essor du numérique rejoignent ce qui était au centre de l'approche philologique traditionnelle.

Nous illustrerons quelques aspects de ces mutations récentes en nous appuyant sur un corpus numérique à l'usage des linguistes médiévistes, la Base de français médiéval (BFM[2]). Il s'agit d'un corpus numérique déjà relativement ancien (initié en 1989) s'appuyant sur une pratique numérique en évolution constante, composé d'éditions de référence (éditions originales et éditions imprimées numérisées), encodées au format XML-TEI et enrichies à de multiples niveaux (métadonnées textuelles, codage interne, segmentation en mots et annotation linguistique). Nous donnerons un aperçu des possibilités nouvelles offertes à l'analyse par ce corpus outillé (section 1), qui motivent la mise en place d'une double chaîne, philologique pour la constitution et la préparation des données textuelles (section 2) et analytique pour leur exploitation outillée (section 3). À travers l'exemple de la BFM, nous tenterons de dégager les contraintes et apports d'un tel cadre méthodologique dans une perspective plus large et plus communautaire.

## 1 Nouvelles avancées méthodologiques dans le domaine de la linguistique diachronique de corpus

Depuis sa création, la Base de français médiéval a été conçue comme un outil dédié à l'étude linguistique historique et diachronique du français. Actuellement exploitée par une communauté internationale de 400 utilisateurs environ, elle a depuis ses origines été le support de nombreuses thèses et travaux de recherche portant sur la langue médiévale. Elle est également utilisée de manière constante par l'équipe en charge de son développement au sein du laboratoire ICAR et de

[1] Les auteurs remercient le LABEX ASLAN (ANR-10-LABX-0081) de l'Université de Lyon pour son soutien financier dans le cadre du programme « Investissements d'Avenir » (ANR-11-IDEX-0007) de l'État Français géré par l'Agence Nationale de la Recherche (ANR).

[2] Le site du projet BFM (http://bfm.ens-lyon.fr) présente la base dans son état actuel et ses objectifs de recherche.

l'ENS de Lyon. Les travaux de recherche menés dans ce cadre alimentent et infléchissent les évolutions de la base. Bien qu'elles portent sur des sujets très variés (de l'évolution de la ponctuation médiévale, de la sémantique des démonstratifs, de l'oral représenté ou des incises en français, pour ne citer que les plus récentes), ces recherches ont pour caractéristique commune de s'appuyer toujours sur les méthodologies définies dans le cadre de la linguistique de corpus. Elles motivent et dirigent l'implémentation dans la base de ressources textuelles et logicielles dont le développement s'effectue de manière parallèle et très étroitement inter-reliée (définition de métadonnées textuelles qui s'articulent aux fonctionnalités de création de corpus/sous-corpus et de contrastes, modèles/outils d'annotation et textes annotés, etc.).

Les ressources numériques ainsi produites permettent de développer des analyses basées sur une démarche empirique, fondée sur des données authentiques et quantifiables, dont les résultats sont reproductibles et vérifiables. Les outils utilisés par l'équipe, qui relèvent de l'approche dite « textométrique » (Lebart & Salem 1994, http://textometrie.ens-lyon.fr), permettent l'analyse quantitative des phénomènes étudiés sans jamais disjoindre les données de leur contexte d'occurrence et des éléments nécessaires à l'interprétation qualitative des résultats.

Une étude récente (Guillot *et al*. 2015) portant sur les caractéristiques de l'oral représenté a permis, par exemple, d'utiliser le calcul statistique de l'Analyse Factorielle des Correspondances (AFC) pour mettre en évidence les spécificités très fortes, stables et durables qui caractérisent le discours direct au Moyen Âge.

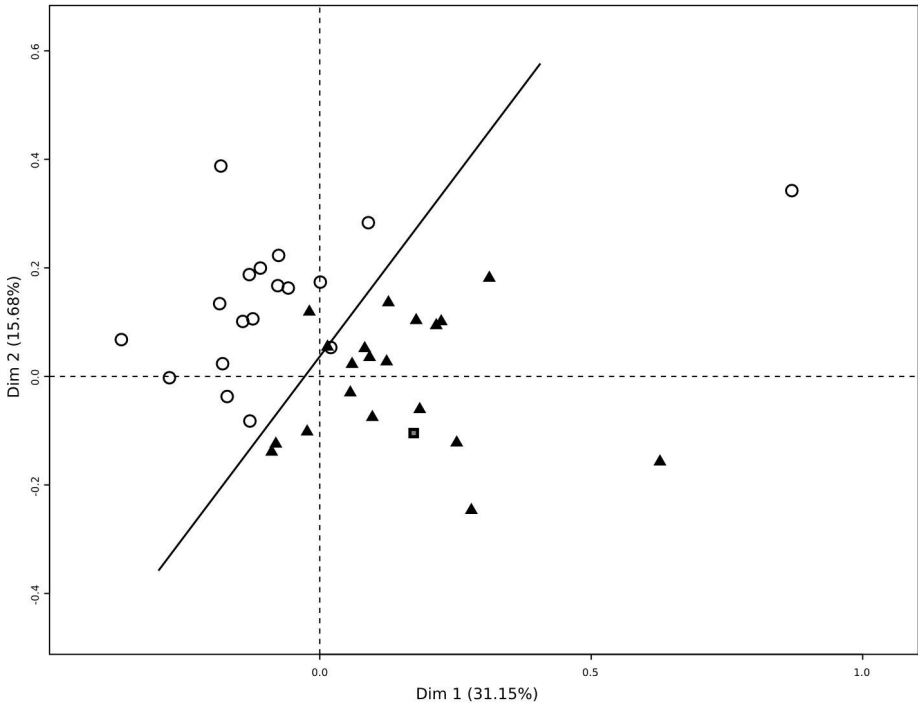


*Figure 1 : AFC des textes du corpus, en distinguant les parties narratives vs au discours direct. Pour le calcul, les textes sont modélisés par la fréquence des catégories morphosyntaxiques qu'ils utilisent. Les parties au discours direct sont représentées par des cercles, et les parties narratives par des triangles.*

Pour cette étude, le balisage numérique des segments au discours direct dans tous les textes de la base a permis de réaliser un calcul d'AFC comparant les fréquences des étiquettes morphosyntaxiques associées aux mots du discours direct à celles des autres parties de chaque texte pour produire une visualisation graphique à deux dimensions positionnant chaque plan de texte (discours direct / parties narratives de chaque texte) en fonction des différences observées. Le plan factoriel montre clairement que l'opposition discours direct / parties narratives constitue un contraste dominant à l'intérieur des textes de la base, puisque les cercles (parties au discours direct) et les triangles (parties narratives) se positionnent d'eux-mêmes de part et d'autre de l'espace

correspondant à cette opposition quels que soient les textes. Nous avons tracé une diagonale séparatrice pour mettre en évidence cette distribution. Le retour aux données qui sont à l'origine de la construction du graphique permet d'interpréter la position originale des parties au discours direct du *Comput* de Philippe de Thaon (représenté par le cercle situé en haut à droite de la fig. 1). Cette position est liée à l'usage très particulier des guillemets dans ce texte : ils n'indiquent pas les segments au discours direct mais servent à citer des mots isolés. Les guillemets étant les marques formelles sur lesquelles a reposé l'encodage du discours direct et sa dissociation des autres parties de textes dans la base, l'usage déviant de ces marques dans le *Comput* explique son positionnement excentrique dans le graphique[3].

L'analyse des étiquettes morphosyntaxiques qui fondent la position relative de chaque point selon le premier axe[4] permet de se faire une idée assez précise des éléments les plus spécifiques au discours direct ou aux autres parties de textes. La fréquence élevée des pronoms personnels, conjonctions de subordination, négations, pronoms impersonnels, interjections et adverbes caractérise le discours direct, celle des noms propres, articles définis, contractions de l'article et des prépositions (*du*, *au*, etc.), noms communs, déterminants cardinaux et participes présents distingue tout ce qui lui est extérieur.

Les études qui sont menées dans un tel cadre d'analyse reposent sur des ressources riches et adaptées. Elles supposent la possibilité d'exploiter les données textuelles avec des outils numériques de synthèse et de recherche. Elles impliquent le traitement d'un volume de données suffisant et d'une diversité assez représentative pour qu'on parvienne à des résultats stables et généraux. Elles imposent aussi un équipement numérique relativement poussé des textes : encodage du discours direct, étiquetage morphosyntaxique de tous les mots, description des unités textuelles grâce à un système de métadonnées permettant l'interprétation des résultats et l'étude de la variation.

Une telle méthode de recherche, qui s'élabore dans un cadre de plus en plus expérimental, vise également à permettre à la communauté scientifique de reproduire les mêmes analyses, grâce à la diffusion, la documentation et la pérennisation des ressources. Cette méthode favorise l'enrichissement continu des textes numériques au fil des analyses linguistiques, les informations rendues disponibles par l'analyse ayant vocation à être associées aux données elles-mêmes pour pouvoir être réutilisées lors de recherches ultérieures. Dans le cas de l'étude citée ci-dessus par exemple, l'analyse des spécificités de l'oral représenté dans la Base de français médiéval amène à réinterpréter la fonction des guillemets dans le texte du *Comput* et à revoir le balisage des séquences au discours direct dans ce texte.

L'amélioration continue des données et le coût inhérent à la préparation et à l'équipement numérique des textes rendent par ailleurs de plus en plus nécessaire le développement partagé et communautaire des ressources. Aux plans juridique et pratique, il est devenu indispensable de permettre la libre circulation et la rediffusion responsable de ces ressources parmi l'ensemble des partenaires qui participent, pour une part variable et à différents niveaux, à leur production et leur exploitation. Le respect de normes et standards internationaux de représentation numérique des textes, l'utilisation de licences de (re-)diffusion ouvertes compatibles avec les juridictions et la jurisprudence internationales sont les principaux instruments de cette politique d'échanges communautaires. Nous essaierons de montrer, dans la suite de cet article, les implications très concrètes de cette méthode de travail concernant les aspects philologiques des textes (section 2) comme les outils d'analyse qui permettent de les exploiter (section 3).

---

[3] Le balisage du discours direct a été réalisé de manière semi-automatique dans tous les textes de la base. Il s'appuie sur les marques graphiques (guillemets ouvrants et fermants) insérées par les éditeurs modernes des textes médiévaux. Les limites de ce balisage automatique sont évidentes. Il permet néanmoins de dégager des tendances générales grâce à l'analyse d'un grand nombre de textes.

[4] Pour des raisons de lisibilité nous avons choisi de ne pas faire figurer ces étiquettes sur le graphique, mais l'outil donne directement accès aux informations qui sous-tendent la position des points du graphique.

## 2 Chaîne philologique ouverte pour l'établissement et l'annotation des textes de la BFM

### 2.1 Principes méthodologiques

Les recherches qui sont menées dans le cadre de la linguistique de corpus et que nous avons illustrées ci-dessus par l'étude des spécificités du discours direct en français médiéval exploitent le plus souvent un volume important de données textuelles. Elles permettent surtout d'appliquer les outils informatiques à l'analyse de données de type et de niveau très variés. Certaines de ces informations concernent les unités textuelles dans leur ensemble (métadonnées textuelles), d'autres sont internes aux textes ou à des parties de textes (structures textuelles, comme les passages au discours direct, les groupes de mots correspondant à des unités inférieures, les mots du texte, les caractères, etc.). Le standard de balisage XML-TEI[5] permet d'encoder toutes ces informations aux niveaux qui leur correspondent (en-têtes pour les métadonnées textuelles, corps du texte pour tout le reste).

La méthodologie de corpus implique par conséquent que l'on identifie et traite séparément au moins trois types d'information : (i) les métadonnées textuelles, qui servent à caractériser les textes, à les regrouper ou à les dissocier, à interpréter les variations observées grâce à l'approche contrastive ; (ii) les structures internes ou unités linguistiques sur lesquelles portent les analyses et qui demandent à être clairement délimitées et balisées (dans l'exemple cité, les segments au discours direct, les limites de mots) ; (iii) les propriétés associées à ces unités linguistiques destinées à être mobilisées lors de l'analyse (étiquettes morphosyntaxiques, lemmes, annotations syntaxiques, etc.). C'est de la combinaison de ces informations multiples que dépendent l'interprétation des résultats et la richesse des analyses.

Lorsqu'il s'agit de corpus de textes médiévaux (ou plus généralement de ceux dont l'édition demande un travail philologique important), des questions méthodologiques supplémentaires doivent par ailleurs être résolues ou en tout cas prises en compte. Il convient en premier lieu de distinguer les sources primaires (les manuscrits, pour l'époque médiévale) des sources secondaires (éditions scientifiques). Il n'est pas envisageable de constituer de grands corpus de textes médiévaux directement à partir des sources primaires, car une bonne transcription de manuscrit est un travail philologique très laborieux qui comprend notamment l'étude de la tradition manuscrite et le choix d'un manuscrit de base, l'identification des erreurs scribales éventuelles, la résolution des nombreuses ambiguïtés des graphies médiévales (telles que les séries de jambages, agglutinations ou abréviations non univoques) et éventuellement la consultation d'autres manuscrits de la même œuvre afin d'éclaircir les passages difficiles. Un tel investissement est discutable si une bonne édition scientifique existe déjà pour un texte. Mais l'utilisation des éditions scientifiques comme source de données pour les corpus numériques pose, d'un autre côté, des problèmes importants.

### 2.2 Ré-ingénierie numérique d'éditions scientifiques existantes

On observe d'abord que les pratiques d'établissement du texte varient considérablement d'une tradition philologique à l'autre, qu'elles évoluent avec le temps et dépendent dans une mesure non négligeable des choix personnels de l'éditeur. Même si dans le domaine de l'édition de textes en français médiéval la tradition « bédiériste »[6] qui consiste à respecter autant que possible le manuscrit de base est largement dominante, le degré de « liberté » que les philologues se donnent dans la correction du manuscrit de base est très variable. Ainsi, M. Plouzeau (1994) a démontré que la dernière version de l'édition de la *Mort Artu* par J. Frappier (1964) ne constituait plus une source de données linguistiques fiable.

Certains aspects de l'établissement de texte sont laissés entièrement à l'appréciation de l'éditeur scientifique. Il s'agit en particulier de la ponctuation et de la segmentation des locutions qui se sont figées et sont devenues des lexies uniques au cours de l'évolution de la langue (comme

[5] http://www.tei-c.org

[6] Nommée ainsi en l'honneur de Joseph Bédier qui en a déclaré les principes dans son étude de la tradition manuscrite du *Lai de l'ombre* (1928).

par exemple la locution prépositionnelle *par mi*, le groupe adverbial *ja mais* ou le syntagme nominal *bon heur*)[7]. Cette hétérogénéité des pratiques pose des problèmes évidents pour l'annotation morphosyntaxique et la lemmatisation du corpus, ainsi que pour les recherches et l'analyse des données textuelles.

Une solution partielle aux problèmes posés par la diversité des pratiques philologiques repose sur la normalisation de l'encodage des textes grâce à l'application des recommandations du consortium TEI (*Text Encoding Initiative*,). On peut ainsi neutraliser les différentes manières d'indiquer les mêmes types d'interventions éditoriales. Par exemple, les fragments restitués par l'éditeur scientifique à la place des lacunes peuvent être signalés, selon les éditions, par des crochets ou par des chevrons (plus rarement). La balise TEI <supplied> peut être utilisée dans les deux cas. Un autre exemple concerne l'indication des passages au discours direct. C'est toujours l'éditeur scientifique qui place les guillemets, car les manuscrits médiévaux n'utilisaient pas cette marque graphique dans cette fonction. En revanche, selon les traditions philologiques et les règles typographiques adoptées dans différents pays, les guillemets peuvent être français (« ») ou anglais (“ ”), être ou ne pas être fermés devant les incises ou entre les prises de parole dans les dialogues. La balise TEI <q> permet d'harmoniser toutes ces pratiques hétérogènes. Enfin, le balisage des mots du texte peut permettre de dissocier la segmentation visuelle réalisée à l'aide des blancs typographiques de la segmentation analytique utilisée dans l'annotation linguistique et dans les requêtes appliquées au corpus. On peut ainsi procéder à une normalisation massive tout en respectant les choix de l'éditeur dans la présentation graphique. L'harmonisation de la segmentation graphique étant cependant une tâche particulièrement lourde, elle n'a pas encore été réalisée dans le corpus de la BFM[8].

Une seconde source de difficulté est liée au fait que l'état de la propriété intellectuelle de nombreuses éditions n'est pas clair. En France (à la différence de l'Allemagne et de l'Italie, par exemple), il n'existe pas de texte législatif spécifique concernant les éditions critiques (Margoni / Perry 2011). Si on considère ces éditions comme des œuvres originales créées par les éditeurs scientifiques, les droits patrimoniaux restent protégés pendant 70 ans après la mort de cet éditeur. Certaines maisons d'édition prétendent détenir les droits de diffusion numérique des textes dont les éditeurs scientifiques sont décédés depuis plusieurs décennies. La recherche d'éventuels ayants droit de ces éditions s'avère souvent très longue et complexe. Les contrats d'édition récents prévoient généralement la cession exclusive des droits de diffusion numérique de toutes sortes à la maison d'édition, ce qui les rend inutilisables dans des corpus numériques, pour lesquels la libre diffusion des données est vitale (Guerreau 2015). Même si la maison d'édition donne son accord pour l'intégration de « son » texte dans un corpus, elle peut le retirer à tout moment, ce qui risque de nuire à la reproductibilité et à la continuité des recherches basées sur ces données. Pour ne plus faire courir ce risque à ses utilisateurs, la Base de français médiéval a été contrainte de retirer un certain nombre de textes (plus d'un million d'occurrences mots au total) en août 2014 suite à la rupture d'une convention avec une maison d'édition.

Selon un autre point de vue, défendu récemment par l'une des parties dans un procès opposant deux maisons d'éditions, le « corps » du texte d'une édition scientifique (à l'exclusion de l'introduction, des notes, des variantes et des annexes de toutes sortes) n'est pas une création de l'éditeur scientifique au sens où l'entend le Code de la propriété intellectuelle et n'est donc pas protégeable. Un jugement de première instance a confirmé cette position, mais la controverse est loin d'être close dans ce débat juridique. Par ailleurs, les notes du texte peuvent comporter des informations très importantes et nécessaires à son analyse (comme l'indication de variantes ou la justification d'une correction).

L'objectif de la Base de français médiéval étant d'offrir à la communauté des chercheurs la ressource la plus riche et la plus fiable possible pour étudier la langue française des premiers textes

[7] Le processus inverse est possible, mais beaucoup plus rare : par exemple le préfixe *tres-* (du latin *trans-*) devenu l'adverbe *très*.

[8] En effet, les annotations syntaxiques réalisées sur certains textes de la BFM dans le cadre du projet SRCMF (Stein & Prévost 2013) reposent sur la segmentation actuelle des mots. Or, toute modification des choix de segmentation lexicale suppose de réaligner ces annotations sur les nouvelles unités lexicales qui pourraient être créées.

à la fin du XV[e] siècle, de multiples facteurs sont pris en compte lors de la sélection des textes à intégrer au corpus. Les œuvres sont d'abord sélectionnées en fonction de leur intérêt linguistique (on cherche à équilibrer le corpus sur le plan diachronique en tenant compte des genres et domaines textuels). La qualité philologique des éditions et leur statut juridique (qui peuvent être facteurs d'exclusion) sont ensuite évalués. En cas d'éditions récentes dont les fichiers de saisie sous un logiciel de traitement de texte sont disponibles et dont les auteurs n'ont pas cédé l'exclusivité des droits à une maison d'édition, la BFM négocie directement avec les éditeurs scientifiques pour obtenir ces fichiers sources et pouvoir les diffuser sous une licence libre. Des éditions plus anciennes sont numérisées aux frais de l'équipe de la BFM, avec l'accord des ayants droit, lorsqu'on les trouve.

## 2.3 Création d'éditions numériques originales

Tous les problèmes liés à la réutilisation d'éditions traditionnelles peuvent être résolus dans des éditions « nativement numériques ». Il est possible, notamment, de fournir plusieurs niveaux de transcription dont chacun est adapté à des usages et à des catégories de lecteurs différents (Guillot *et al*. à par. 2015 ; Marchello-Nizia *et al*. à par. 2015).

Dans la pratique, il nous semble qu'une représentation à deux niveaux, qu'on peut dénommer « normalisée » et « diplomatique » peut satisfaire la grande majorité des utilisateurs. Le niveau normalisé se rapproche de la tradition de l'édition des textes littéraires, avec toutefois l'application de règles plus explicites concernant notamment la ponctuation, la segmentation graphique des mots et la résolution des abréviations. Le niveau diplomatique se rapproche davantage du système graphique du document source : les lettres restituées à la place des abréviations sont signalées par des italiques, les distinctions « ramistes » (phonétiques) des lettres *i/i* et *u/v* ne sont pas introduites, les diacritiques modernes ne sont pas ajoutés et aucune marque de ponctuation n'est utilisée, lorsqu'il n'y en a pas dans le document transcrit. La segmentation graphique correspond dans la mesure du possible à celle du document source[9]. Ce type de transcription peut être indispensable pour certains types de recherche linguistique, en particulier dans le domaine de la morphologie (Schøsler 2004 : 463). Pour réaliser une transcription « à deux niveaux », il n'est pas nécessaire de transcrire deux fois le texte source. Il suffit d'utiliser un petit nombre de raccourcis typographiques, dans le cadre d'une convention de transcription utilisant un mécanisme de caractères spéciaux, qui permettent de générer automatiquement les deux types de transcription à partir d'un fichier unique. Par exemple, le caractère dièse permet de signaler dans *#Dieu* que la majuscule du nom propre est due à la normalisation éditoriale et que la graphie du document source comporte une minuscule.

Les principes de la segmentation lexicale pour les outils d'annotation linguistique et pour le moteur de recherche peuvent être clairement définis et appliqués dans le cadre de grandes collections d'éditions numériques et, idéalement, partagés par la communauté internationale des philologues. Il convient de souligner que la normalisation de la segmentation au niveau du codage informatique n'empêche pas l'éditeur scientifique d'appliquer ses propres choix de segmentation visuelle dans l'édition à l'écran ou imprimée. En règle générale, le codage de la segmentation la plus fine est préférable, car il est plus simple de regrouper que de découper des unités *a posteriori*.

Le modèle économique des éditions numériques diffère considérablement des éditions imprimées. Le coût de la fabrication et de la diffusion du livre est important et peut justifier la cession des droits à l'éditeur commercial. Pour les éditions numériques basées sur une chaîne de production bien réglée et disposant d'une plateforme de diffusion adaptée, c'est le travail philologique de l'éditeur scientifique qui représente l'investissement principal. Le coût d'hébergement d'une ressource sur le web est relativement faible, et les services à forte valeur ajoutée (impression à la demande, export dans un format particulier) peuvent être proposés aux lecteurs. Ceci rend tout-à-fait possible la diffusion des éditions numériques sous une licence libre de type *Creative Commons* ou similaire. Cela est important non seulement pour faciliter l'accès à la lecture de ces éditions par les membres de la communauté académique et un public plus large, mais

[9] Dans certains cas, faute d'instrument de mesure précis, la lecture et la décision de transcrire un blanc entre deux mots du manuscrit reste à l'appréciation de l'éditeur.

aussi et surtout pour permettre leur intégration dans des archives ouvertes, dans des corpus divers et variés, ainsi que dans le web de données. La possibilité d'accéder aux données primaires des travaux de recherche pour reproduire leurs résultats est un élément important de leur scientificité. Enfin, plus une ressource numérique est utilisée et reproduite, plus il y a de chances qu'elle puisse s'adapter aux évolutions technologiques constantes.

La diffusion ouverte des données implique l'utilisation de formats de représentation ouverts, le respect des normes et standards d'encodage et la documentation des pratiques particulières à une équipe. Pour ce qui concerne l'encodage d'éditions scientifiques numériques, le cadre proposé par le consortium TEI (déjà évoqué plus haut) semble à ce jour le mieux adapté. Les avantages de la TEI sont sa riche expérience (plus de 25 ans), la variété des types de textes et d'éditions pris en charge, la souplesse des schémas de balisage, sa documentation extensive et sa communauté active. Certains des points forts de la TEI peuvent également devenir ses faiblesses. Le très grand nombre de balises disponibles pour l'encodage et le fait qu'il existe toujours plusieurs façons de faire pour encoder un même phénomène rendent difficile la mise au point d'outils d'analyse. La documentation fournie par la TEI ne suffit pas toujours pour expliciter les choix faits au niveau d'un projet de recherche particulier. Pour cette raison, la TEI recommande de personnaliser le schéma de balisage utilisé par un projet ou par une communauté et fournit un mécanisme facilitant cette personnalisation et sa documentation. La BFM utilise le balisage TEI pour ses éditions numérisées depuis le début des années 2000 et documente ses pratiques de manière précise (Bertrand *et al*. 2013). Les éditions nativement numériques appliquent le même schéma de base que le reste de la BFM, mais utilisent un certain nombre de balises supplémentaires, notamment pour la représentation des transcriptions multi-niveaux.

## 3 Chaîne analytique ouverte pour l'exploitation textométrique des textes de la BFM

L'application d'un sous-ensemble précis des recommandations de la TEI documenté dans le cadre de la BFM a non seulement permis de mettre en place un réseau d'échanges de textes entre partenaires partageant les mêmes pratiques philologiques, mais il a également permis à ce corpus de textes d'être intégré dans la plateforme TXM pour son analyse et sa diffusion.

Développée initialement dans le cadre d'un projet financé par l'ANR en 2007-2010, cette plateforme a pour objectif de pérenniser et de mutualiser les développements informatiques d'outils textométriques comme Hyperbase, Lexico 3, Le Trameur, DTM et Weblex. La textométrie est une méthode d'analyse de corpus textuels développée depuis les années 1980[10] combinant des outils statistiques appliqués aux différentes unités des textes (analyse factorielle, calcul de spécificités, classification, analyse de cooccurrents) et des outils documentaires (listes de mots, recherche plein texte de patrons de mots pour l'établissement de concordances, lecture des éditions de textes du corpus). Son implémentation dans la plateforme TXM a été l'occasion d'élargir la méthode aux corpus textuels richement encodés en XML-TEI et annotés par différents outils de traitement automatique de la langue (comme le lemmatiseur TreeTagger) et de produire une version pour poste Windows, Mac OS X ou Linux (appelée « logiciel TXM ») ainsi qu'une version serveur pour l'accès par Internet (appelée « portail TXM »), les deux versions partageant la même plateforme de base pour l'exploitation des corpus.

La mutualisation de la construction et de la maintenance de la plateforme est obtenue par un mode de développement ouvert appelé « open-source » bien établi dans les projets de recherche en informatique depuis 20 ans, qui fonctionne sur deux plans. D'une part tout partenaire peut accéder aux sources du logiciel pour l'adapter ou l'améliorer en respectant les termes de la licence de diffusion[11]. Et d'autre part, la plateforme intègre elle-même de nombreux composants logiciels développés par d'autres projets open-source. En particulier l'environnement de calcul statistique

[10] Voir http://textometrie.ens-lyon.fr/spip.php?rubrique80.

[11] La licence GNU GPL V3 : http://www.gnu.org/licenses/gpl-3.0.fr.html.

R[12], le moteur de recherche CQP[13] et la plateforme Eclipse[14] pour la version pour poste de TXM.

Le fait de pouvoir accéder aux sources du logiciel TXM est par ailleurs un gage de scientificité des travaux réalisés grâce à cet outil, parce qu'il ne fonctionne pas comme une boîte noire. Tous ses calculs sont décomposables et vérifiables à partir de ses sources. Le fait de déléguer certains calculs à d'autres composants open-source permet de profiter de leurs performances et de leurs améliorations constantes par leur communauté de développement. Mais il faut s'assurer que chaque composant soit bien maintenu par une communauté de développeurs dynamique, par des institutions ou des entreprises au risque qu'il ne soit un jour plus développé et ne puisse plus suivre les évolutions technologiques et continuer à fonctionner. Auquel cas on doit soit le remplacer par un composant open-source équivalent soit le maintenir soi-même.

La pérennisation du développement repose sur deux plans : d'une part l'utilisation d'un langage de programmation correspondant à un standard industriel reconnu et développé selon un mode communautaire ouvert (Java[15]) et une architecture logicielle standard (OSGi[16]) d'autre part l'utilisation d'une plateforme de versionnage des sources du logiciel, qui permet la traçabilité de l'attribution et de la datation de toute modification apportée aux sources et offre la possibilité de revenir à une version antérieure de n'importe quelle date.

Conçue dès l'origine comme devant être capable d'exploiter des corpus textuels richement encodés en XML-TEI et annotés finement au niveau des mots, la plateforme TXM a pu utiliser la BFM comme corpus de validation de ses capacités d'intégration et d'exploitation de corpus textuels riches en encodage et annotations.

La chaîne analytique de la BFM commence par un processus d'importation des fichiers sources encodés en XML-TEI dans la plateforme TXM à l'aide d'un module d'importation de sources appelé « XML-TEI BFM ». Ce module a été développé spécialement pour ce corpus à partir de la documentation des pratiques d'encodage XML-TEI des textes de la BFM telle qu'elle est publiée sur le site du projet de la Base. Il est chargé d'interpréter les fichiers source de sorte à construire le « modèle de corpus » exploité par TXM. Les métadonnées nécessaires et utiles à l'analyse des textes sont extraites des entêtes TEI, les éléments TEI pertinents pour l'analyse (comme par exemple les éléments <q> contenant le discours direct) sont indexés et certaines informations sont projetées au niveau des unités lexicales afin de simplifier les requêtes de recherche. D'autres éléments (comme les notes éditoriales) sont exclus de la surface du texte afin de ne pas être mélangés avec le corps du texte dans les recherches et les décomptes mais sont intégrés aux éditions pour aider à la lecture des textes. Les éditions sont paginées en fonction des sauts de page encodés dans les sources numérisées. Une fois importée dans TXM à l'aide de ce module d'importation, la BFM bénéficie de tous les services d'analyse offerts par la plateforme dans sa version pour poste ou dans sa version portail. Le portail BFM (http://txm.bfm-corpus.org) est un portail TXM hébergeant le corpus BFM. Il offre des services supplémentaires par rapport à la version pour poste de personnalisation de pages d'accueil ou de documentation, de création de comptes utilisateurs et de contrôle d'accès texte par texte.

## Conclusion : une synergie entre les chaînes philologique et analytique pour une ressource libre

Aujourd'hui la BFM est consultée et analysée au moyen d'un logiciel libre (la plateforme TXM) et offre un accès libre aux sources de ses textes. Ces sources sont établies par une chaîne philologique complète et ouverte, de façon analogue à la chaîne d'analyse qui repose sur le logiciel TXM, lui-même développé en open-source. L'emboîtement entre ces deux chaînes est rendu possible par un usage précis du standard de représentation des textes XML-TEI. Développée à l'origine pour l'échange de représentations numériques de textes entre partenaires, la TEI

12 http://www.r-project.org.
13 http://cwb.sourceforge.net.
14 https://eclipse.org.
15 https://www.jcp.org.
16 http://www.osgi.org.

commence donc à mettre en relation des projets d'établissement de corpus de textes avec des projets de développement d'outils d'analyse et d'exploitation qui relèvent pourtant souvent de communautés de recherche très différentes en termes d'objectifs et de mode de fonctionnement. L'adoption parallèle d'un mode de fonctionnement ouvert par les deux chaînes pour faciliter la mutualisation et la traçabilité des développements (établissement de texte d'un côté, implémentation de méthode de l'autre) nous semble être un gage de pérennité et de scientificité des travaux pouvant être réalisés à l'aide de la BFM.

## Références


Bédier, Joseph (1928), « 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, Lauranne, Lavrentiev Alexei, Pincemin, Bénédicte, Guillot, Céline, Heiden, Serge et Lascar, Justine (2014), *Tutoriel TXM pour la BFM*. Version 2.0, Lyon, ENS de Lyon. <http://txm.bfm-corpus.org/files/Tutoriel_TXM_BFM_V1.pdf>.

Frappier Jean (éd.) (1964), *La mort Artu*, Genève, 3e éd.,Paris : Droz, Minard.

Guerreau, Alain (2015), *L'avenir de la philologie textes anciens et domaine public*. <halshs-01112213>

Guillot, Céline, Lavrentiev, Alexei , Rainsford, Thomas , Marchello-Nizia, Christiane et Heiden, Serge (à par. 2015), « La ‹ philologie numérique › : tentative de définition d'un nouvel objet éditorial » in Buchi, Éva, Chauveau, Jean-Paul et Pierrel, Jean-Marie (éd.), *Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013)*, 3 volumes. Strasbourg : Société de linguistique romane/ÉLiPhi. <halshs-00846767>

Guillot, Céline, Heiden, Serge, Lavrentiev, Alexei et Pincemin, Bénédicte (2015), « L'oral représenté dans un corpus de français médiéval (9e-15e) : approche contrastive et outillée de la variation diasystémique », in Jeppesen Kragh, Kirsten et Lindschouw, Jan (éds), *Les variations diasystémiques et leurs interdépendances dans les langues romanes -Actes du Colloque DIA II à Copenhague (19-21 nov. 2012)*, Strasbourg : Éditions de linguistique et de philologie, p. 15-28 <halshs-00760647>.

Lebart, Ludovic, Salem, André (1994), *Statistique Textuelle*, Paris : Dunod.

Marchello-Nizia, Christiane, Lavrentiev, Alexei, Guillot-Barbance, Céline (à par. 2015), « Édition électronique de la *Queste del saint Graal* », in Trotter, David (éd.), *Manuel de la philologie de l'édition*, Berlin, Boston : De Gruyter Mouton.

Margoni, Thomas et Perry, Mark (2011), « Scientific and Critical Editions of Public Domain Works : An Example of European Copyright Law (Dis)Harmonization », *Canadian Intellectual Property Review*, Vol. 27, 2011, p. 157. <http://ssrn.com/abstract=1961535>.

Plouzeau, May (1994), « A propos de *La Mort Artu* de Jean Frappier », *Travaux de linguistique et de philologie,* vol. 32, 1994, p. 207-221.

Schøsler, Lene (2004) « Historical corpora. Problems and Methods », in Bozzi, Andrea, Cignoli, Laura et Lebrave, Jean-Louis (éds), *Digital technology and philological disciplines*, *Linguistica computazionale*, vol. XX-XXI, Pisa, Roma : Istituti editoriali e poligrafici internazionali, 2004, p. 455-472.

Stein, Achim et Prévost, Sophie (2013), « Syntactic annotation of medieval texts: the Syntactic Reference Corpus of Medieval French (SRCMF) », in Bennett, Paul, Durrell, Martin, Scheible, Silke et Whitt,Richard (éds), *New Methods in Historical Corpora, Corpus Linguistics and International Perspectives on Language, CLIP,* Vol. 3, p. 275-282. Tübingen: Narr.