CORPUS

Corpus

25 | 2024
La constitution de corpus en diachronie longue

La BFM 2022 : un corpus pour les recherches diachroniques en français médiéval et au-delà

BFM2022: a corpus for diachronic research on Medieval French and beyond

Alexei Lavrentiev and Céline Guillot-Barbance

OpenEdition Journals

Electronic version
URL: https://journals.openedition.org/corpus/8601
DOI: 10.4000/corpus.8601
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Electronic reference
Alexei Lavrentiev and Céline Guillot-Barbance, "La BFM 2022 : un corpus pour les recherches diachroniques en français médiéval et au-delà", *Corpus* [Online], 25 | 2024, Online since 25 January 2024, connection on 01 February 2024. URL: http://journals.openedition.org/corpus/8601 ; DOI: https://doi.org/10.4000/corpus.8601

La BFM 2022 : un corpus pour les recherches diachroniques en français médiéval et au-delà

BFM2022: a corpus for diachronic research on Medieval French and beyond

Alexei Lavrentiev and Céline Guillot-Barbance

Introduction

1 La Base de français médiéval (BFM) fait partie des corpus diachroniques français à la plus longue histoire (33 ans !) et les plus importants pour la période linguistique ancienne. Fondée à la fin des années 1980, cette base a toujours été développée à l'ENS de Lyon (anciennement ENS Fontenay/Saint-Cloud puis ENS Lettres et Sciences humaines), sous l'égide d'abord de C. Marchello-Nizia, puis de C. Guillot-Barbance et d'A. Lavrentiev, avec le concours de S. Heiden pour le versant outil[1]. Elle est actuellement administrée à l'UMR 5317 IHRIM et hébergée dans l'infrastructure nationale Huma-Num (http://txm.bfm-corpus.org). Elle est librement interrogeable en ligne ou sur poste grâce à la plateforme d'analyse textuelle TXM[2] (Heiden *et al.* 2010a) et téléchargeable dans l'entrepôt NAKALA[3].

2 Le projet initial de la BFM était de diffuser à la communauté scientifique un corpus d'ancien français (12^{e}-13^{e} s.), complémentaire à celui du *Dictionnaire du moyen français* (DMF ; 1330-1500) et au corpus Frantext (postérieur à 1500) du *Trésor de la langue française* (TLF). Comme ces derniers, elle était composée d'éditions scientifiques de textes intégraux imprimés et numérisés. Seul le corps du texte était traité et le renvoi aux numéros de pages et/ou de vers permettait de faire le lien avec les éditions papier et l'apparat critique correspondant. Dans sa version première, la BFM était diffusée par des concordances statiques accessibles seulement sur CD-ROM. Avec les moyens très limités qui étaient les siens, elle a ainsi pu faire partie de la première génération des corpus numériques du français et a dès le départ été développée en même temps que l'outil permettant de l'exploiter (le logiciel ANALYSER de P. Bonnefois).

3 Le début des années 2000 a marqué le tournant numérique de la Base, lorsqu'elle est devenue accessible en ligne grâce au logiciel Weblex développé par S. Heiden à l'ENS Lettres et Sciences humaines de Lyon. Ce choix s'est avéré décisif d'un triple point de vue : (i) la diffusion et la notoriété de la BFM se sont largement accrues ; (ii) le logiciel Weblex a considérablement étendu ses possibilités d'interrogation en rendant disponibles les outils développés par l'école lexicométrique de l'ENS (cooccurrences, segments répétés, spécificités, lexicogrammes, etc.) en plus des traditionnelles concordances et listes de fréquences ; (iii) sous l'influence de S. Heiden puis d'A. Lavrentiev, les normes et standards internationaux XML-TEI favorisant l'ouverture et l'échange des données ont été rapidement adoptés et adaptés (Heiden & Barbance-Guillot 2003, Heiden *et al.* 2010b). Trois des principales forces actuelles de la BFM (ouverture, prise en compte de la représentativité des données, interopérabilité) découlent plus ou moins directement de ce choix[4]. La représentativité des données fera l'objet de la section 1 du présent article, l'interopérabilité sera abordée dans la section 2.

4 La mise en ligne du corpus à partir de 2000 et du site du projet (http://bfm.ens-lyon.fr) à partir de 2005 ont eu une autre conséquence quelques années plus tard, avec la mise en demeure de la Base par la librairie Droz. L'action juridique initiée par la maison d'édition a entraîné la fermeture de la Base pendant trois ans (2007-2010) et a convaincu l'équipe de s'engager définitivement dans une politique d'ouverture maximale (Guillot *et al.* 2017). Peu après sa réouverture, la BFM est devenue accessible en ligne grâce à un portail TXM et elle diffuse désormais ses textes sous licence ouverte Etalab[5] et les apparats critiques sous licence CC BY-NC-SA 3.0 FR[6]. Depuis 2012, six mises à jour ont donné naissance à six corpus identifiés par l'année de leur publication : BFM2012 (3 300 000 mots), BFM2013 (4 700 000 mots), BFM2014 (3 550 000 mots), BFM2016 (4 100 000 mots), BFM2019 (4 700 000 mots), BFM2022 (6 400 000 mots). Durant ces années et sous l'impulsion des programmes de recherche qui se sont enchaînés[7], la BFM s'est beaucoup diversifiée du point de vue textuel, elle s'est enrichie de multiples couches d'annotation et s'est étendue du point de vue diachronique, beaucoup de textes de très ancien français et de moyen français s'étant ajoutés. Le corpus BFM2022 est actuellement utilisé par une communauté d'environ 400 chercheurs, étudiants et membres du grand public intéressés par la langue, la littérature, l'histoire et la culture médiévales.

5 Le développement parallèle d'une Base textuelle et d'un outil de recherche associé (Weblex puis TXM) ont beaucoup compté dans les orientations et la dynamique collective du projet. On essaiera de montrer en quoi cette synergie permet de traiter la question de la représentativité du corpus à la fois du point de vue des données et des méthodologies de recherche (section 1). On insistera également sur la nécessaire articulation entre, d'une part, la standardisation et l'interopérabilité des données sur le plan technique (section 2), et, d'autre part, la création de corpus diachroniques étendus pour des recherches en diachronie longue (section 3).

1. Représentativité et équilibrage du corpus

6 La représentativité d'un corpus est primordiale pour permettre la généralisation des résultats obtenus à partir de ses données. Idéalement, un corpus destiné à l'étude des états anciens d'une langue devrait être aussi exhaustif que possible, car les textes sont

la seule source de données disponible. Dans la réalité, l'exhaustivité est difficile à atteindre avec les moyens techniques et les ressources humaines dont la linguistique diachronique dispose actuellement. L'exhaustivité ne signifie par ailleurs pas une représentativité parfaite, car les manuscrits et les imprimés qui nous sont parvenus ne constituent qu'une infime partie de la production langagière. Le plus important est donc de disposer d'une modélisation adéquate des paramètres de variation pertinents et de pouvoir décrire les textes du corpus selon ces paramètres, afin de savoir précisément ce que ce corpus représente et quelles sont les « zones d'ombre » non couvertes ou sous-représentées. Cette connaissance doit guider le développement d'un corpus qui se veut représentatif, dans la limite des contraintes qui s'imposent. Parmi ces contraintes on peut citer l'existence même des sources primaires (très peu de textes français ont été composés avant le milieu du 12e s., il existe très peu de textes en prose avant la fin du 12e s. ou peu de textes littéraires non anglo-normands à la même époque), mais aussi la disponibilité et la qualité des éditions pouvant être intégrées au corpus. La question du coût de la numérisation, de l'annotation et de la maintenance des ressources est également loin d'être anodine. Les projets financés, qui sont devenus depuis une vingtaine d'années incontournables pour développer la recherche, ont chacun des objectifs précis et amènent en général à privilégier une catégorie de textes ou une autre.

7 Pendant la première phase du développement de la BFM, qui s'est déroulée pratiquement sans financement spécifique, l'objectif était de numériser les textes les plus « symboliques » de la littérature française médiévale (comme la *Chanson de Roland* ou la *Quête du saint Graal*), mais le corpus s'enrichissait aussi grâce à des échanges ponctuels ou à des contributions de doctorants qui travaillaient sur un texte particulier. À la différence des corpus échantillonnés, comme la partie « littéraire » du Corpus d'Amsterdam (Dees 1987), les œuvres ont toujours été prises intégralement, dans la lignée de l'école « néo-firthienne » de la linguistique de corpus (McEnery & Hardie 2012 : 152).

8 La question de la représentativité de la BFM s'est posée dans les années 2000, à l'occasion de l'ouverture de son site web. Progressivement, un système de « descripteurs » (ou paramètres de variation) a été élaboré afin d'évaluer la couverture de la Base et de définir les priorités de son développement jusqu'à présent.

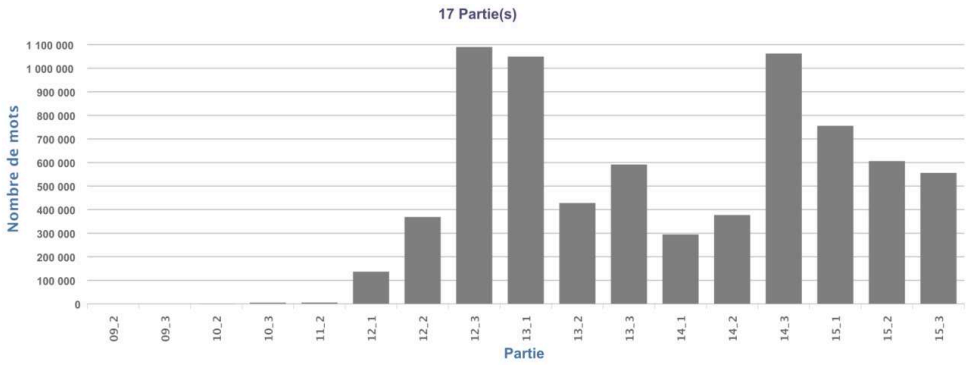
1.1. Vue d'ensemble de la BFM2022

9 Comme la BFM2022 est un corpus diachronique, le premier paramètre de variation à prendre en compte est chronologique. Pour la majorité des textes médiévaux, il existe un décalage plus ou moins important entre la date supposée de composition d'une œuvre et la date d'exécution du manuscrit de base de l'édition intégrée au corpus. Par ailleurs, les deux dates ne sont souvent connues qu'approximativement, à quelques dizaines d'années près. La BFM permet aux utilisateurs de travailler avec les deux dates, mais privilégie la date de composition pour l'organisation et la description du corpus. Un système de normalisation des dates approximatives a été mis en place afin de permettre le tri et la sélection des textes par les utilisateurs.

10 Le corpus BFM2022 paraît assez équilibré lorsque ses dimensions sont observées par siècle (entre 1,6 et 2 millions de mots, du 12e au 14e s.). En revanche, en choisissant une granularité plus fine (33 ans, soit début, milieu et fin de chaque siècle) on constate que

certaines périodes sont sous-représentées (Figure 1). Si pour le début et le milieu du 12e s. cette sous-représentation est inévitable en raison du nombre limité de textes qui nous sont parvenus, le « creux » au milieu du 13e s., ainsi qu'au début et au milieu du 14e s., est un déséquilibre que nous chercherons à réduire dans les prochaines versions de la BFM. La sur-représentation de la fin du 12e siècle s'explique par les apports d'un projet ANR (CoRPTeF) visant à créer un corpus aussi exhaustif que possible pour le très ancien français, tandis que pour la fin du 14e s. un seul « texte »[8] très volumineux, le *Registre criminel du châtelet*, pèse près de 37 % de la masse totale.

Figure 1. Dimensions du corpus BFM2022 par tiers de siècle (début / milieu / fin)

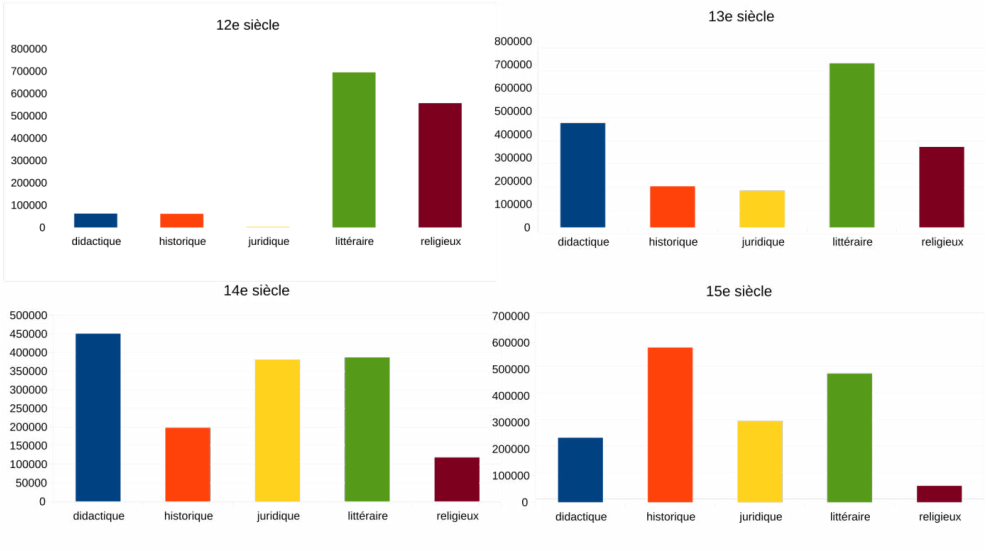


11 Le second paramètre de variation primordial en ancien français est celui de la variation dialectale. Rares sont toutefois les textes qui ne présentent des traits que d'un seul dialecte. À partir du 13e s., la part des textes sans « coloration » dialectale devient de plus en plus importante. L'attribution d'un dialecte à un texte repose dans la BFM sur les fiches bibliographiques du *Dictionnaire étymologique de l'ancien français* (DEAFBiblEl)[9] et, dans certains cas, sur les introductions des éditions. L'étiquette « non défini » est utilisée lorsque les informations dont nous disposons ne permettent pas de déterminer le dialecte dominant. Au 12e siècle, le dialecte anglo-normand est nettement dominant (25 % du corpus). Viennent ensuite le champenois (16 %), le wallon (14 %), le normand (9,5 %) et le picard (8,9 %). Au 13e s. la part des textes sans dialecte défini atteint 37 %, elle est légèrement inférieure à celle du picard (38 %). La part du lorrain, qui se place en troisième position, ne dépasse pas les 7,5 %. Au 14e s. les textes dont le dialecte n'est pas défini ou est étiqueté « parisien » dominent largement (73 % du corpus), tandis qu'au 15e s. la part du picard remonte à nouveau (28 %, contre 32 % pour les textes sans dialecte défini). Cette évolution qui peut sembler étonnante s'explique encore une fois par le biais provoqué par certains textes très volumineux : le *Registre criminel du Châtelet* au 14e s. (parisien) et les *Chroniques* de Jean Froissart au 15e s. (picard, 13 % du volume).

12 Une autre dimension (ou plutôt un ensemble de dimensions) de variation concerne le contenu et la structure interne des textes. Dans les métadonnées de la BFM, nous distinguons la forme du texte (vers, prose ou mixte) de son domaine fonctionnel (didactique, historique, juridique, littéraire ou religieux) et du genre proprement dit. Les genres sont assez nombreux et évoluent fortement au cours des siècles, ce qui rend la tâche d'équilibrage quasiment impossible sur un empan diachronique important. Les domaines sont en revanche plus stables et leur nombre est délimité, même si au fil des siècles de nouveaux domaines s'ouvrent aux textes vernaculaires, tandis que d'autres perdent en importance. Les dimensions des domaines siècle par siècle sont présentées dans la Figure 2.

13 Le domaine littéraire est bien représenté dès la fin du 12e s. et jusqu'à la fin du 15e. Sa part baisse progressivement de 50 % à 25-29 % aux 14e et 15e s., dans une répartition plus équilibrée. Le domaine religieux est très important au 12e s. (40 %) et diminue fortement, jusqu'à 4 % au 15e s. Les domaines didactique et historique augmentent au contraire. Le didactique est surreprésenté au 14e s. (29 %), en raison de la présence dans le corpus de plusieurs textes relativement volumineux (le *Mesnagier de Paris*, le *Livre de Ethiques d'Aristote* de Nicole Oresme, *De la Erudition* de Jean Daudin, etc.). Au 15e s., c'est le domaine historique qui occupe une part importante (35 %), ce qui s'explique par la présence des *Chroniques* de Jean Froissart déjà mentionnées et des *Mémoires* de Philippe de Commynes. Le domaine juridique est pratiquement absent jusqu'au milieu du 13e s., puisque les chartes françaises n'apparaissent qu'à cette époque (hormis quelques rares cas antérieurs).

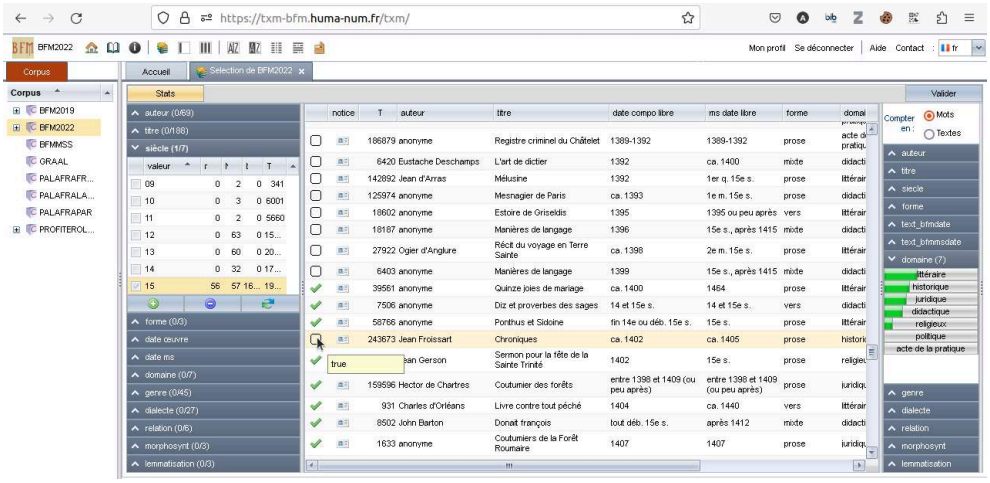
Figure 2. Dimensions des domaines par siècle dans le corpus BFM2022



1.2. Stratégies et outils d'équilibrage du corpus

14 Lors de l'analyse quantitative du corpus, plusieurs méthodes existent afin de réduire les biais produits par les déséquilibres. La technique la plus simple consiste à créer des sous-corpus par sélection de textes et à réduire ainsi les parties sur-représentées. Une interface spécifique de « sélection de textes » permet de visualiser dynamiquement les dimensions du sous-corpus selon les critères choisis (Figure 3).

Figure 3. Interface de sélection de textes de la BFM. Dans le volet à gauche on utilise des critères pour la création d'un sous-corpus (15e s. dans l'exemple), dans la partie centrale on ajoute ou supprime des textes individuellement (dans l'exemple on décoche les Chroniques de Froissart) et dans le volet droit on visualise les dimensions du sous-corpus par critère (ici par domaine).



15 Il est également possible de recourir à l'échantillonnage des textes très longs, mais dans ce cas, il est nécessaire de récupérer les fichiers XML TEI TXM du corpus (disponibles dans l'entrepôt Gitlab d'Huma-Num[10]), pour éditer manuellement ou à l'aide d'un script XSLT les textes à échantillonner avant de réimporter le corpus dans TXM. Un tutoriel consacré à la création de corpus personnalisés à partir des textes de la BFM est en cours de rédaction et sera publié sur le site de la BFM prochainement. À moyen terme, la fonctionnalité de création de corpus échantillonnés pourra être ajoutée au portail BFM.

16 Un autre moyen de pallier les déséquilibres du corpus consiste à utiliser des méthodes statistiques appropriées. Le score de spécificités que le portail BFM permet de calculer sur une partition[11] implémente la méthode proposée par P. Lafon (1980). Pour chaque partie considérée alors comme un sous-corpus, le calcul tient compte de sa taille, de la taille du corpus entier, et du nombre d'occurrences de chaque forme, lemme ou étiquette morphosyntaxique dans ces deux ensembles. Le score indique le logarithme décimal de la probabilité que le nombre d'occurrences soit aussi élevé (ou bas)[12] dans la partie en cas de distribution aléatoire. Le score situé entre -2 et +2 (au moins une chance sur 100) est généralement considéré comme « banal » et jugé non significatif. Si une partie est très petite ou si le nombre d'occurrences dans le corpus entier est très bas, le score de spécificités n'est jamais très élevé. En revanche, un score élevé (positif ou négatif) est un indice fiable de sur- ou sous-représentation (Lebart *et al.* 2019 : 123).

2. Standardisation et interopérabilité des données et des outils

17 L'ouverture maximale des données et le recours aux outils d'analyse et de publication libres et ouverts ont toujours été le souhait de l'équipe de la BFM. Si l'incertitude sur les droits d'auteur applicables aux éditions critiques de textes médiévaux a empêché pendant des années la diffusion libre des ressources, depuis le jugement en appel dans le procès Droz contre Garnier, l'appartenance du « corps du texte » médiéval au domaine public (quel que soit l'éditeur scientifique ou commercial) ne fait guère de doute, en tout cas en France (Mouron 2017).

18 Cela rend possible la mise à disposition sous licence ouverte[13] des ressources construites sur la base de ces textes, à savoir dans notre cas des documents XML balisés selon les recommandations de la *Text Encoding Initiative*. Ce système de balisage est un gage de pérennité et d'interopérabilité des données. Il inclut la structuration des textes (divisions, paragraphes, groupes de vers…), un système de références (numérotation des pages et des vers selon l'édition de référence) et l'ensemble des métadonnées associées à chaque document (teiHeader). L'apparat critique est exclu, sauf si un accord a été conclu avec les ayants-droit. Le discours direct est balisé dans la quasi-totalité des textes de la BFM, ce qui permet d'accéder dans une certaine mesure à l'étude de l'oralité médiévale.

19 La politique de standardisation porte également sur les annotations linguistiques, dont une part croissante est vérifiée par des spécialistes (9,5 % du corpus pour les lemmes et étiquettes, 6,1 % pour les étiquettes seules)[14]. Le jeu d'étiquettes morphosyntaxiques original de la BFM conçu par Sophie Prévost est nommé « Cattex ». Il est particulièrement adapté au français médiéval mais convient également à l'ensemble des langues romanes. Toutefois, dans le but de faciliter l'échange de données et la création de corpus multilingues et étendus sur une diachronie longue, la conversion au format défini par le projet *Universal Dependencies* (UD)[15] a été effectuée en 2019. Les étiquettes « ud-pos » et « ud-feats » sont disponibles dans les corpus BFM2019 et BFM2022. La conversion est entièrement automatique et n'entraîne pas de perte de données. Certaines adaptations du standard UD ont néanmoins dû être réalisées. Ainsi, la distinction entre les verbes « pleins » et les auxiliaires au niveau des parties du discours, inexistante dans Cattex, n'a pas été introduite. En cas de contraction, la double étiquette a été maintenue (par exemple, PRE.DETdef devient ADP.DET pour les contractions d'une préposition avec le déterminant défini), bien que le standard UD préconise l'annotation des formes « profondes » à l'origine de la contraction. Il convient également de noter que l'étiquetage de la BFM se limite au niveau « minimal » de Cattex (catégorie et sous-catégorie éventuelle), tous les traits morphologiques prévus par UD ne sont donc pas renseignés.

20 Dans le domaine de la lemmatisation, la standardisation est plus difficile en raison de l'absence d'un référentiel de lemmes unique et des variations de segmentation inhérentes aux évolutions diachroniques et aux pratiques hétérogènes des éditeurs scientifiques. Certains projets comme le *Dictionnaire électronique de Chrétien de Troyes*[16] se réfèrent au dictionnaire Tobler & Lommatzsch (1925-1995), d'autres utilisent les lemmes du *Dictionnaire du moyen français* (DMF)[17]. La BFM a fait le choix de privilégier les lemmes du DMF, en raison notamment de leur compatibilité avec les entrées du TLFi pour le français moderne, ce qui facilite la création de corpus diachroniques longs. La vérification des lemmes s'accompagne de la normalisation de la segmentation lexicale selon les principes éditoriaux de la BFM (Lavrentiev *et al.* 2021). Afin d'établir des correspondances entre lemmes provenant de sources multiples, l'équipe de la BFM a compilé un lexique morphologique ouvert FROLEX à partir du lexique LGeRM de l'ATILF, du lexique AFRLEX d'Achim Stein et de ses propres textes lemmatisés. Ce lexique, accompagné d'un tableau de correspondance entre les référentiels des lemmes, est mis à disposition de la communauté sur la plateforme Github[18].

21 L'interopérabilité de toutes ces données annotées permet de les utiliser pour l'entraînement d'outils d'annotation automatique diffusés indépendamment de la Base. Un modèle linguistique pour le logiciel TreeTagger (Schmid 1994)[19] est disponible sur le

site de la BFM, ainsi que sur le site du TreeTagger. Les données de la BFM ont également permis d'entraîner un outil de nouvelle génération, le RNNTagger (Schmid 2019)[20]. Ce dernier est plus performant, mais demande plus d'espace disque et un processeur plus puissant.

22 Enfin, l'effort de standardisation porte non seulement sur l'annotation du corps des textes de la BFM, mais aussi sur leurs métadonnées. Les noms des auteurs, lorsqu'ils sont connus, sont alignés avec le référentiel IDREF, des liens entre le catalogue de la BFM et les notices bibliographiques du DEAF sont établis dans la mesure du possible. La BFM participe par ailleurs à l'élaboration d'un thésaurus de typologie textuelle initiée dans le cadre du Consortium CAHIER (Marasescu-Galleron *et al.* 2022), sur lequel seront à terme alignés tous les termes caractérisant les genres et les domaines des textes.

3. La BFM et les corpus de recherche en diachronie longue

23 Grâce à sa politique de standardisation et d'échange de textes, la Base de français médiéval a une longue expérience dans la constitution de corpus en diachronie longue. Dès les années 2002-2005, un échange avec l'ATILF avait permis l'intégration réciproque de textes d'ancien et de moyen français dans le Grand Frantext et le corpus BFM. Plus récemment, le projet DEMOCRAT a permis la création et l'annotation en chaînes de référence d'un corpus français allant du 12e au 21e siècle (voir la présentation du projet dans ce numéro). Pour illustrer davantage la diversité des situations et des défis rencontrés, nous avons choisi de présenter plus précisément deux exemples.

24 Le projet ANR/DFG *Passage du latin au français* (PaLaFra, 2015-2018)[21] portait sur l'étude de la transition latino-romane et visait pour cela à constituer et à exploiter un corpus bilingue (latin tardif, 6e-8e s., 350 000 mots / ancien français, 9e-13e s., 1 million de mots). Ce corpus a agrégé deux types de sources (*Monumenta Germaniae Historica* et BFM) dans une même plateforme (TXM). La sélection des données s'est appuyée sur le système de métadonnées de la BFM (Guillot & Lavrentiev 2007). Elle s'est essentiellement fondée sur des critères pragmatiques, en fonction du profil conceptionnel des textes (Koch & Oesterreicher 2011) dans le but de rendre compte de la « scripta latina rustica » (Sabatini 1968) et de la polynormativité des textes latins. Les textes français ont été choisis selon les mêmes principes, en favorisant les genres textuels représentés en latin. Il s'agissait ainsi de privilégier l'accès aux variations préfigurant l'émergence du français et de permettre l'étude contrastive des deux langues. Le corpus bilingue offre la possibilité d'études longitudinales portant sur des vies de saints, chartes et textes historiques, tout en garantissant une certaine diversité dans chaque langue (ajout de formulaires et lettres en latin, de textes littéraires, didactiques et d'autres genres religieux en français). Le caractère multilingue du projet a fait avancer le chantier de standardisation des étiquettes morphosyntaxiques évoqué dans la section précédente. Les corpus PALAFRALAT-V2-0 et PALAFRAFRO-V2-2 sont accessibles sur le portail de la BFM et sous forme de corpus « binaires » destinés à la version pour poste de TXM.

25 La *Grande grammaire historique du français* (GGHF) publiée en 2020 chez De Gruyter (Marchello-Nizia *et al.* 2020), constitue un ouvrage de référence pour l'histoire du français des origines à aujourd'hui (9e-20e s.). L'une des originalités du projet était de

s'appuyer sur un corpus diachronique spécifique. Les sources rassemblées dans ce corpus sont multiples (BFM pour la période ancienne, *Bibliothèques virtuelles humanistes* pour le 16e siècle, DMF et Frantext à partir du 16e siècle). Les métadonnées sélectionnées reposent sur le système mis en place dans la BFM (auteur / titre / date / siècle / forme / domaine / genre / dialecte) et l'outil d'exploitation est la plateforme TXM. Ce projet a été pour nous l'occasion de développer l'échantillonnage des textes dans le corpus dit « noyau échantillonné » (2 055 891 mots), qui se distingue du corpus « intégral » (13,5 millions de mots)[22] par le fait qu'il est équilibré, mieux contrôlé et plus maniable (Prévost 2020 : 41).

Conclusion et perspectives

26 Depuis ses origines la Base de français médiéval constitue un corpus textuel diachronique et ses limites se sont progressivement étendues au fil des décennies. Dans son développement, elle a dû faire face au défi de la diversification et de l'équilibrage des données pour garantir leur représentativité. Elle s'est également enrichie grâce à une politique d'échange et, pour ce faire, elle a adopté dès le début des années 2000 les standards numériques pérennes. Ces standards ont à leur tour permis d'enrichir les textes grâce à un formatage très fin et à une annotation multi-niveau. La plateforme d'analyse TXM développée en parallèle de la Base a été conçue pour tirer parti de cette richesse de formatage et d'annotation, tout en offrant les apports des outils de l'analyse textométrique.

27 Les projets de recherche dans lesquels l'équipe s'est impliquée depuis une quinzaine d'années ont également permis de se former à la création de corpus divers et d'étendue très variable. La BFM n'ayant pas vocation à s'élargir à l'infini du point de vue diachronique, il s'agit pour nous à présent d'aider nos utilisateurs à rassembler eux-mêmes les données utiles pour leurs recherches diachroniques dans des corpus spécifiques. Un premier pas dans cette direction a été la création d'un « kit » de corpus diachronique pour TXM accompagné d'une documentation technique. La mise en place dans TXM de différents types d'outils d'annotation (au fil du texte, en concordance, par macro) est aussi une manière de développer des stratégies relativement souples, simples et rapides, pour faire face à l'hétérogénéité des données en diachronie longue. Grâce à ses ressources textuelles et aux outils qui la diffusent, la Base de français médiéval s'efforce ainsi de favoriser et de contribuer au développement de la recherche diachronique sur le français.

BIBLIOGRAPHY

Dees A. (1987). *Atlas des formes linguistiques des textes littéraires de l'ancien français*. Tübingen : Max Niemeyer Verlag.

Guillot C. & Lavrentiev A. (2007). *Manuel de description de textes pour la Base de Français Médiéval*, v. 1.2, Lyon, Projet BFM, http://ccfm.ens-lsh.fr/IMG/pdf/Manuel_Descripteurs_BFM_v1.2.pdf.

Guillot C., Heiden S. & Lavrentiev A. (2017). « Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique », *Diachroniques* 7 : 168-184.

Heiden S. & Guillot C. (2003). « Capitalisation des savoirs par le web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval », P. Kunstmann, F. Martineau et D. Forget (éd.), *Ancien et moyen français sur le Web : enjeux méthodologiques et analyse du discours [Actes du colloque d'Ottawa, 4-5 oct. 2002]*. Ottawa : Éditions David, 77-92.

Heiden S., Magué J.-P. & Pincemin B. (2010a). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », S. Bolasco, I. Chiari et L. Giuliano (éd.), *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*. Rome, Italie : Edizioni Universitarie di Lettere Economia Diritto, 1021-1032, https://halshs.archives-ouvertes.fr/halshs-00549779.

Heiden S., Guillot C., Lavrentiev A. & Bertrand L. (2010b). *Manuel d'encodage XML-TEI des textes de la Base de Français Médiéval*, v. 4.0, Lyon, Projet BFM, http://bfm.ens-lyon.fr/IMG/pdf/Manuel_Encodage_TEI.pdf.

Lafon P. (1980). « Sur la variabilité de la fréquence des formes dans un corpus », *Mots. Les langages du politique* 1 : 127-165.

Lavrentiev A., Guillot C. & Heiden S. (2021). « Enjeux philologiques, linguistiques et informatiques de la philologie numérique : l'exemple de la segmentation des mots », *Diachroniques. Revue de Linguistique française diachronique* 8 : 77-102.

Lebart L., Pincemin B. & Poudat C. (2019). *Analyse des données textuelles*. Québec : Presses de l'Université du Québec.

Koch P. & Oesterreicher W. (2011). *Gesprochene Sprache in der Romania. Französisch, Italienisch, Spanisch*. Berlin : De Gruyter.

Marasescu-Galleron I., Idmhand F., Lavrentiev A., Demonet M.-L. & Réach-Ngô A. (2022). « Décrire un corpus d'auteurs », F. Idmhand et I. Marasescu-Galleron (dir.), *Dix ans de corpus d'auteurs*. Paris : Éditions des archives contemporaines, 31-46. DOI : 10.17184/eac.5806.

Marchello-Nizia C., Combettes B., Prévost S. & Scheer T. (éd.) (2020). *Grande Grammaire Historique du Français*. Berlin : De Gruyter Mouton.

McEnery T. & Hardie A. (2012). *Corpus linguistics method, theory and practice*. Cambridge : Cambridge University Press.

Mouron P. (2017). « La restitution d'un manuscrit du Moyen Âge n'est pas une œuvre originale », *Revue Lamy Droit de l'Immatériel* 143 : 1-5.

Prévost S. (2020). « Une grammaire fondée sur un corpus numérique », C. Marchello-Nizia, B. Combettes, S. Prévost & T. Scheer (éd.), *Grande Grammaire Historique du Français*. Berlin : De Gruyter Mouton, 37-53.

Sabatini F. (1968). « Dalla scripta latina rustica alle scripte romanze », *Studi medievali* 9 : 320-358.

Schmid H. (1994). « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK, https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Schmid H. (2019). « Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts », *DATeCH*, mai 2019, Brussels, Belgium, https://www.cis.uni-muenchen.de/~schmid/papers/Datech2019.pdf.

Tobler A. & Lommatzsch Erhard F. (1925-1995). *Altfranzösisches Wörterbuch*. Mayence, Wiesbaden, Stuttgart, F. Steiner, 11 vol.

NOTES

1. Font également partie de l'équipe M. Decorde (développement informatique), N. Pontal (éditorialisation et accompagnement juridique) et A. Collignon (encodage XML TEI).
2. https://txm.gitpages.huma-num.fr/textometrie/Presentation.
3. https://nakala.fr/collection/10.34847/nkl.93ee3ts1.
4. Les autres traits distinctifs de la BFM sont l'enrichissement multi-niveau des données (métadonnées, lemmes, étiquettes morphosyntaxiques, catégories syntaxiques (cf. la présentation du projet PROFITEROLE dans ce numéro), annotation des chaînes de référence (cf. la présentation du projet DEMOCRAT dans ce numéro), encodage du discours direct) et la production d'éditions numériques originales (notamment l'édition de la Queste del saint Graal par C. Marchello-Nizia et A. Lavrentiev).
5. https://www.etalab.gouv.fr/licence-ouverte-open-licence
6. https://creativecommons.org/licenses/by-nc-sa/3.0/fr
7. La BFM n'ayant pas de financement propre récurrent, elle se développe depuis une quinzaine d'années grâce à des programmes de recherche locaux, nationaux et internationaux (dont 2 projets ANR/DFG et 6 projets ANR). La liste complète des projets est accessible en ligne : http://bfm.ens-lyon.fr/spip.php?article339.
8. Nous mettons le mot entre guillemets, car il s'agit en réalité d'un ensemble d'actes, donc de textes indépendants rassemblés dans deux volumes, et deux fichiers numériques, pour des raisons pratiques.
9. https://alma.hadw-bw.de/deafbibl
10. https://gitlab.huma-num.fr/bfm/bfm-textes-diffusion
11. Les fonctionnalités d'analyse statistique sont généralement plus avancées dans le logiciel TXM pour poste que sur le portail BFM. Pour les spécificités, le logiciel pour poste permet d'effectuer cette analyse sur un sous-corpus et non seulement sur une partition. Il est également possible d'éditer la table de données (« table lexicale ») avant de lancer le calcul.
12. Par convention le score positif caractérise les fréquences supérieures à la moyenne et le score négatif les fréquences inférieures.
13. Conformément à la politique de diffusion de données adoptée à l'ENS de Lyon, la BFM utilise la licence ouverte Etalab (https://www.etalab.gouv.fr/licence-ouverte-open-licence).
14. Toutes les annotations, y compris les lemmes et les étiquettes morphosyntaxiques, sont diffusées dans les mêmes conditions que le corps du texte.
15. https://universaldependencies.org
16. http://www.atilf.fr/dect
17. http://www.atilf.fr/dmf
18. https://github.com/sheiden/Medieval-French-Language-Toolkit
19. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger
20. https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger
21. Le projet réunissait l'ENS de Lyon et l'Université de Lille du côté français, et les Universités de Regensburg et de Tübingen du côté allemand : http://www.palafra.org.

22. Ce corpus est composé à son tour d'un corpus noyau non échantillonné (9 millions de mots) et du corpus « complémentaire » (4,5 millions de mots).

ABSTRACTS

The *Base de Français Médiéval* (BFM) is one of the oldest corpora of Medieval French (9th-15th centuries), and it is one of the most used by diachronic linguists and more broadly by all those interested in the history of French. It is the result of a collaboration between linguists, philologists and specialists in the textometric method implemented in the TXM platform. This article presents characteristics of the BFM2022 corpus focusing on the representativeness and interoperability of the data. It illustrates how digital tools can be used for data balancing and introduces the multi-level standardization strategy (textual structure, references, metadata, annotations) which allowed the Base to be integrated into more extensive diachronic corpora for long diachronic research.

La Base de français médiéval (BFM) fait partie des corpus de français médiéval (9e-15e s.) les plus anciens et les plus utilisés par les linguistes diachroniciens et plus largement par tous ceux qui s'intéressent à l'histoire du français. Elle est le fruit d'une collaboration entre linguistes-philologues et spécialistes de la méthode textométrique implémentée dans la plateforme TXM. L'article présente un état des lieux du corpus BFM2022 focalisé sur la représentativité et l'interopérabilité des données. Il illustre l'apport des outils numériques pour l'équilibrage des données et présente la stratégie de standardisation multi-niveau (structure textuelle, références, métadonnées, annotations) ayant permis à la Base de s'intégrer à des corpus diachroniques plus étendus pour des recherches en diachronie longue.

INDEX

Mots-clés: Français médiéval, textométrie, représentativité, interopérabilité, standardisation, diachronie longue
Keywords: Medieval French, textometry, representativeness, interoperability, standardization, long diachrony

AUTHORS

ALEXEI LAVRENTIEV

CNRS, IHRIM – UMR 5317

CÉLINE GUILLOT-BARBANCE

ENS de Lyon, IHRIM – UMR 5317