HAL
open science

Base de français médiéval et transcriptions de manuscrits: recherche de complémentarité

Alexei Lavrentiev

To cite this version:

Alexei Lavrentiev. Base de français médiéval et transcriptions de manuscrits: recherche de complémentarité. XXIVe Congrès International de Linguistique et de Philologie Romanes, Aug 2004, Aberystwyth, Royaume-Uni. pp.405-410. halshs-02869691

HAL Id: halshs-02869691

https://shs.hal.science/halshs-02869691v1

Submitted on 16 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Alexei Lavrentiev

Base de français médiéval et transcriptions de manuscrits: recherche de complémentarité

Les bases de données textuelles sont devenues des outils incontournables pour toute recherche sur la langue, en diachronie en particulier. Certaines, comme la Base de Français Médiéval (BFM),[1] sont fondées sur des éditions critiques, tandis que d'autres, comme le projet *Charrette*,[2] représentent directement les manuscrits sous la forme de transcriptions diplomatiques plus ou moins détaillées.[3]

Les éditions critiques présentent l'avantage d'être faciles à numériser et permettent de constituer un corpus de taille importante en un temps relativement bref, la transcription détaillée de manuscrits étant longue et coûteuse. Si l'on compare deux projets de bases de textes en français médiéval, dont la durée et le coût sont assez proches, la différence dans la taille des corpus obtenus apparaît clairement. Ainsi, la BFM a accumulé en dix ans 74 textes (2 850 000 occurrences-mots), tandis que le projet *Charrette* se limite à la tradition manuscrite d'un seul roman (204 000 occurrences-mots dans les 8 manuscrits).

La question se pose cependant de savoir si les éditions critiques sont une source de données suffisamment ‹authentique› et donc fiable pour des études linguistiques.

P. Bourgain et F. Vielliard, dans leur ouvrage récent (2002: 14-20), rappellent la distinction traditionnelle entre les deux grands paradigmes méthodologiques concernant l'établissement du texte pour une édition critique. La première, dite «lachmanienne»,[4] consiste à tenter de reconstituer le texte de l'original perdu en étudiant les filiations et les rapports des manuscrits et en éliminant les «fautes communes» aux différentes branches. L'autre méthode, dite «bédiériste», consiste à choisir un «bon» manuscrit et à le reproduire fidèlement, en limitant les corrections aux «cas d'extrême et presque évidente nécessité» (Bédier 1928). Pour les textes de français médiéval, la plupart des éditeurs suivent depuis les années 20 du XXe siècle les consignes de Bédier, en respectant globalement les graphies de leur manuscrit de base. Cette approche est sans aucun doute préférable pour des études linguistiques, car la reconstitution d'un texte n'est jamais libre d'interprétations plus ou moins arbitraires de l'éditeur scientifique.[5] Or, pour le linguiste, les «erreurs» des scribes

---

1 Collection de textes en français médiéval, des origines au début du XVIe siècle; projet initié et dirigé par C. Marchello-Nizia, UMR 5191 ICAR – ENS-LSH, CNRS, http://bfm.ens-lsh.fr.

2 Edition multimédia de la tradition manuscrite du roman du *Chevalier de la Charrette* de Chrétien de Troyes; projet initié et dirigé jusqu'en 2003 par K.D. Uitti, Université de Princeton, http://www.princeton.edu/~lancelot.

3 Ces deux grands types d'approche de la constitution de corpus électroniques de textes anciens ont été analysés plus en détail dans (Heiden / Lavrentiev 2004).

4 Du nom du philologue allemand Karl Lachmann (1793-1851), spécialiste notamment de l'édition des textes de l'Antiquité, au sujet duquel est née la polémique.

5 D'après L. Stanovaïa, «toute segmentation, quelque argumentée qu'elle semble […], détruit l'unité sémantique et linguistique du texte. En plus, la méthodologie qui propose aux chercheurs

sont souvent plus intéressantes que les corrections de l'éditeur, même si ces dernières sont basées sur les données d'autres manuscrits.

Cependant, même l'édition traditionnelle la plus «fidèle» d'un manuscrit ne reproduit jamais exactement le système graphique du témoin transcrit. Les conventions de transcription qui sont généralement admises (résolution des abréviations, suppression des variantes calligraphiques de certaines lettres, rétablissement de certaines distinctions graphiques, ajout de signes diacritiques, normalisation de la segmentation graphique et de la ponctuation)[6] visent à faciliter la lecture et la compréhension du texte. Ces corrections rendent les éditions inutilisables en tant que sources de données pour certains types de recherche (études de systèmes graphiques, de la segmentation et de la ponctuation), mais ne remettent pas en cause, dans l'ensemble, la fiabilité des données pour les recherches lexicologiques, morphologiques et syntaxiques.

Dans les faits, les pratiques éditoriales sont plus complexes et plus variées. Certains textes n'ont pas été réédités depuis le début du XXe, voire la fin du XIXe siècle. A l'époque, une partie des éditeurs scientifiques n'hésitaient pas à corriger largement les «erreurs» des scribes ou des copistes. L'article de M. Plouzeau (1994), qui a mis en évidence les nombreuses interventions de J. Frappier dans le texte du manuscrit de base de son édition de la *Mort Artu* (1964),[7] a semé des doutes sur la fiabilité d'éditions plus récentes, qui étaient jusqu'alors considérées comme fiables et utilisées dans plusieurs projets de corpus informatisés. Faute d'une évaluation précise de la «fidélité» de toutes les éditions à l'origine des corpus informatisées, dans les domaines des formes graphiques, du lexique, de la morphologie et de la syntaxe, le caractère incontestable des résultats de recherches basées sur ces corpus ne peut être assuré.

Il y a donc une contradiction entre la nécessité d'avoir d'une de «gros» corpus pour certaines recherches linguistiques, qui ne peuvent être constitués dans des délais et à un coût raisonnables que par la numérisation d'éditions modernes, et, d'autre part, l'assurance de l'authenticité des données et la richesse de ces données. Il convient de préciser cependant que les recherches qui nécessitent une représentation riche des données manuscrites (études des systèmes graphiques ou de la ponctuation par exemple) n'exigent pas forcément la prise en compte de la totalité des textes, des échantillons de taille limitée à quelques milliers d'occurrences pouvant suffire.

Une solution qui pourrait résoudre à la fois le problème de la fidélité des données présentes dans les gros corpus et de la représentation enrichie des manuscrits sans exiger la mobilisation de ressources financières et humaines trop importantes consisterait, à notre avis, à compléter les bases de textes existantes (telles que la BFM) avec des transcriptions diplomatiques fragmentaires (de 1000 à 2000 occurrences) des manuscrits des textes qui les composent. Le projet de transcription diplomatique «fine» d'échantillons de manuscrits servant de base aux éditions critiques, et à partir de là l'évaluation de la «fidélité» de ces dernières font partie des axes prévus du développement de la Base de Français Médiéval.

---

d'attribuer chaque forme archaïque, littéraire, correcte, etc. à l'auteur, de même que chaque innovation, erreur, régionalisme, etc., au scribe, aboutit aux conclusions toujours subjectives et souvent simplistes» (2003: 244).

6 Cf. (Roques 1926; Bourgain / Vielliard 2002: 61-67).

7 La conclusion de M. Plouzeau est sans équivoque: la langue du manuscrit de base de l'édition de J. Frappier «ne peut pas s'appréhender à travers la seule édition [de 1964]» (1994: 209).

Dans le présent article nous exposons les résultats des premiers essais de «diplomatisation» de la BFM, ainsi que les taux d'«indices de fidélité» des éditions que nous avons obtenus sur les échantillons traités.

Les références des manuscrits et des éditions des textes concernés, ainsi que les abréviations qui les désignent, sont présentées dans le tableau qui figure ci-dessous:

Abréviation	Texte	Edition	Manuscrit	Echantillon
Artois	*Le Roman du Comte d'Artois*	(Seigneuret 1966)	BNF, fr. 11610	f° 1r-4r, 1161 occ.
Beaumanoir	Philippe de Beaumanoir, *Coutumes de Beauvaisis*	(Salmon 1899)	BNF, fr. 11652	f° 4v-5v, 1265 occ.
Froissart	Jean Froissart, *Chroniques*	(Diller 1972)	Vatican, Reg. Lat. 869	f° 1v-2r, 1053 occ.
Monstrelet	Enguerrand de Monstrelet, *Chroniques*	(Douët d'Arcq 1857)	BNF, fr. 2682	f° 233r-234r, 831 occ.
QJM	*Les XV Joyes du Mariage*	(Rychner 1963)	Rouen, BM, 1052	f° 87r-89r, 659 occ.

Les extraits sélectionnés ont été tirés de textes du corpus de la BFM et ensuite corrigés d'après les manuscrits en tenant compte des abréviations, des variantes de lettres, des segmentations particuliers et de la ponctuation. Pour représenter ces traits dans la transcription informatisée, nous avons utilisé le système de balisage en XML élaboré dans le cadre du projet *Charrette* et basé sur les recommandations de la TEI.[8]

Nous avons ensuite collationné les deux versions, éditée et transcrite, de chaque fragment en comparant les graphies des mots, mais aussi la segmentation et la ponctuation. En cas de différence au niveau des mots, nous avons distingué les changements lexicaux (*conte* → *duc*), morphologiques (*vertu* → *vertus*) et grapho-phonétiques (*ozeroit* → *oseroit, cheulx* → *ceulz*), ainsi que l'ajout ou la suppression d'une majuscule initiale (*Messire ame* → *messire Amé*). Dans la segmentation, nous avons relevé les cas de fusion et de séparation (*pour ce* → *pource, que il* → *qu'il, tresforte* → *tres forte*). En ce qui concerne la ponctuation, nous avons repéré les cas de maintien, d'ajout ou de suppression de la ponctuation forte ou faible, ainsi que le «renforcement» ou l'«affaiblissement» de la ponctuation. Des changements syntaxiques, qui peuvent se traduire par la modification de l'ordre des mots ou par la suppression ou l'ajout d'éléments de la phrase, sont plus difficiles à traiter avec cette méthode. Ils peuvent toutefois être repérés au niveau lexical (ajout, suppression ou déplacement d'un mot) et analysés au cas par cas ensuite.

L'ajout d'une majuscule et les changements dans la ponctuation et la segmentation graphique n'affectent pas la «fidélité» de l'édition, car la modernisation de ces éléments du système graphique est explicite dans toutes les éditions. Néanmoins, l'analyse du travail de l'éditeur dans ce domaine peut s'avérer intéressante pour comprendre dans quelle mesure la modernisation du système graphique relève des normes générales ou des préférences personnelles de l'éditeur.

---

[8] Consortium international visant à élaborer les normes pour l'encodage de tous types de bases textuelles (Text Encoding Initiative, http://www.tei-c.org).

L'analyse des extraits dépouillés des textes choisis à montré que les changements au niveau des mots sont extrêmement rares dans *Froissart, QJM* et *Artois*. Aucune modification morphologique des formes du manuscrit n'a été repérée dans ces textes – du moins dans les fragments que nous avons analysés. Les changements étiquetés comme «lexicaux» (1 occ. dans *Artois* et 2 occ. dans *QJM*) sont des corrections de coquilles évidentes: *manie → manière*,[9] *pare → preparé*.[10] Les changements graphiques sont également très peu nombreux (5 occ. dans *Froissart* et 1 occ. dans *Artois*). Il s'agit de 4 cas de remplacement *z → s* (*maiz → mais, eulz → euls, mieulz → mieuls, lez → les*), d'une correction *fors → forts* (sans doute une «faute de frappe» de l'éditeur, car on trouve dans la même édition les graphies *mors, Edouwars*, etc.) et d'une correction de coquille (*uusques → jusques*). De façon générale l'«indice de fidélité» de ces éditions est proche de 99,5% des formes graphiques. Ces éditions peuvent donc servir de source de données pour tous les types de recherches linguistiques qui n'exigent pas la représentation enrichie des graphies manuscrites (ponctuation, segmentation, abréviations, etc.).

Monstrelet et *Beaumanoir* présentent un cas différent. Les deux éditions datent du XIXe siècle. Am. Salmon, qui a édité les *Coutumes de Beauvaisis* de Philippe de Beaumanoir (1899), s'est inspiré de la méthode lachmanienne et s'est efforcé de reconstruire le texte original de l'œuvre, sans choisir un manuscrit de base. Néanmoins, il a privilégié le texte du manuscrit BNF, fr. 11652, qu'il considérait comme l'«un des mss. les plus rapprochés de l'original» (Salmon 1899: XIX), ce qui ne l'a pas empêché d'en corriger massivement la morphologie (38 occ.) et les graphies (106 occ.). L'indice général de la fidélité de cette édition n'atteint que 86%.

L'essentiel des corrections morphologiques porte sur certains pronoms et mots grammaticaux (*aussint → aussi, lui → li, auquant → aucun, les queles → lesqueus*) et sur les formes nominales (ajout de *-s* final au cas sujet singulier de *vertu* et de quelques autres substantifs et adjectifs, rétablissement de la consonne finale dans *bailli(f)*). Dans le domaine phono-graphique, Am. Salmon a remplacé assez régulièrement *z → s* et a effectué de nombreux autres changements de façon plus sporadique (*raisons → resons, apelee → appelee, meesmement → meismement, peresce → perece*). L'éditeur a également introduit les formes élidées du pronom relatif *que* (*que il → qu'il*). Même si beaucoup de corrections sont appuyées par les graphies des autres manuscrits, le texte de l'édition d'Am. Salmon ne peut pas être considéré comme représentatif de la langue des manuscrits au niveau des graphies et de la morphologie. En revanche, les changements lexicaux et syntaxiques y sont assez rares (6 occ.), il est donc possible d'utiliser cette édition pour des recherches linguistiques dans ces domaines.

L'édition de *Monstrelet* effectuée par L. Douët d'Arcq (1857) s'écarte nettement moins de son manuscrit de base. Son indice de fidélité s'élève à 93%. Les corrections effectuées par l'éditeur scientifique portent essentiellement sur les graphies. Les seules exceptions sont une correction morphologique du pronom-sujet de la 3e pers. pl. *jl → ilz* (cette dernière graphie est utilisée dans les 4 occ. de ce pronom dans le ms.) et une correction lexicale (ou plutôt factuelle) *conte de sombreset → duc de Sombreset*. En ce qui concerne les graphies, l'éditeur a régulièrement remplacé le digraphe *ch* (lat. *c + e,i*) par *c* (*chest → c'est, princhе → prince*, etc.) et il a normalisé quelques graphies qui pourraient paraître «exotiques» à un

[9] Contexte: «il pourra trouver manie(re) d'avoir ses plaisirs» (*QJM*)

[10] Contexte: «au logis qui estoit retenu et (pre)paré pour rechepvoir le conte d'Artois» (*Artois*)

lecteur moderne: *heubt* → *eut* (passé simple du verbe *avoir*), *auoecq* → *avec*. En cas de variation graphique dans le manuscrit, l'éditeur a choisi une forme unique: *seigneur / segneur* → *seigneur, engles / angloix* → *Anglois*. L'édition de Douët d'Arcq est en somme assez fidèle à son manuscrit de base pour être utilisée comme source de données pour des recherches linguistiques, à l'exception de celles qui concernent la variation graphique et l'histoire de l'orthographe.

Paradoxalement, ce sont les éditions de *Beaumanoir* et de *Monstrelet* qui respectent le plus/mieux la ponctuation des manuscrits. Dans ces deux textes, on trouve les marques de ponctuation aux mêmes endroits dans le ms. et l'édition dans 48% des cas. Ce même indice varie de 27 à 38% dans les autres textes. Le changement de ponctuation, dont la fréquence est de loin la plus élevée dans toutes les éditions, est l'ajout d'une ponctuation faible.

Ces premiers résultats obtenus à partir de la confrontation des éditions utilisées pour la création de la BFM avec leurs manuscrits de base montrent que les éditions donnent des manuscrits une représentation plus ou moins fidèle selon les niveaux du système linguistique et qu'une évaluation systématique de l'écart entre l'édition et le manuscrit augmenterait nettement la valeur des données de la base. Aussi, la prise en compte de la «fidélité» des éditions n'est qu'une partie du projet du développement de la BFM et de sa transformation en un corpus organisé. Il est bien d'autres problèmes méthodologiques importants qui restent à résoudre, tels que celui de la datation des textes (entre la date supposée de la composition du texte et celle du ms.)[11] ou celui de la typologie externe (genres, domaines, dialectes) et interne (phénomènes linguistiques) des textes de la base.

Bibliographie

Bédier, Joseph (1928): *La tradition manuscrite du Lai de l'Ombre, réflexions sur l'art d'éditer les anciens textes*. In: *Romania* 44, 161-196; 321-356.

Bourgain, Pascale, Françoise Vielliard (2002): *Conseils pour l'édition des textes médiévaux*. Fascicule III. Textes littéraires. Paris: C.T.H.S. – Ecole Nationale des Chartes.

Diller, G. T. (éd.) (1972): Froissart: *Chroniques*. Premier livre. Genève: Droz.

Douët d'Arcq, Louis (éd.) (1857-1862): Enguerrand de Monstrelet: *Chronique* (6 vol.). Paris: Vve de J. Renouard.

Frappier, Jean (éd.) (^{3}1964): *La Mort le Roi Artu*. Genève: Droz.

Heiden, Serge, Céline Guillot (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. In: Kunstmann, Pierre *et al*. (éds.): *Ancien et moyen français sur le Web, enjeux méthodologiques et analyse de discours.* Ottawa: Les éditions David, 77-92.

Heiden, Serge, Alexei Lavrentiev (2004): Ressources électroniques pour l'étude des textes médiévaux: approches et outils. In: *Revue française de linguistique appliquée* IX-1, 99-118.

Kunstmann, Pierre (2000): Ancien et moyen français sur le Web: textes et bases de données. In: *RLiR* 253-254, 17-42.

Plouzeau, May (1994): A propos de *La Mort Artu* de Jean Frappier. In: *TraLiPhi* XXXII, 207-221.

Roques, Mario (1926): Rapport de la 2e commission. Etablissement de règles pratiques pour l'édition des anciens textes français et provençaux. In: *Romania* 52, 243-249.

[11] Cf. Stanovaïa (2003: 243-245) pour la discussion sur la différence entre le *texte* et le *manuscrit*.

Rychner, Jean (éd.) (1963): *Les XV Joyes du Mariage*. Genève: Droz.

Salmon, Amédée (éd.) (1970): Philippe de Beaumanoir. *Coutumes de Beauvaisis.* Texte critique publié avec une introduction. Tome premier. Paris: A. et J. Picard. (Paris 1899)

Seigneuret, Charles (éd.) (1966): *Le Roman du Comte d'Artois*. Genève: Droz.

Stanovaïa, Lydia (2003): La standardisation en ancien français. In: Goyens, Michèle, Werner Verbeke (éds.): *The dawn of the written vernacular in Western Europe*. Leuven: Leuven University Press, 241-272.