Exploitations et valorisations des données numériques connexes à l'édition

Ariane Pinche

Exploitations et valorisations des données numériques connexes à l'édition : le cas des *Seint Confessor* de Wauchier de Denain

Ariane Pinche

Septembre 2021

Introduction

Toute édition est le fruit d'une époque et d'une école philologique [Trotter, 2015, Introduction, pp. 3-6]. Les pratiques les plus répandues aujourd'hui dans la philologie médiévale découlent soit de l'école « lachmanienne » qui cherche à rétablir le *urtext*[1], soit de l'école « bédiériste » qui cherche à établir la transcription du meilleur témoin. La *new philogogy* depuis les années 1990, influencée par les principes de « mouvance » établi par P. Zumthor [Zumthor, 1972] et de « variance » de B. Cerquiglini [Cerquiglini, 1989], invite à non pas établir un texte qui se rapprocherait de celui de l'auteur, notion anachronique pour la période médiévale, mais à s'intéresser à la copie et à ses caractéristiques linguistiques propres. Les différentes approches philologiques ont donné naissance à des pratiques nationales relativement divergentes. Ainsi, la tradition italienne tend vers le neolachmanismes et accorde encore beaucoup de place à la reconstruction textuelle, la plupart des éditions françaises suivent la méthode bédiériste, tandis que les Anglo-saxons composent des éditions documentaires ou diplomatiques [Trotter, 2015, Introduction, pp.1-18]. Les pratiques peuvent également diverger d'un champ disciplinaire à un autre entre historiens, linguistes, littéraires en proposant des degrés de fidélité à la copie différents[2]. Malgré ces divergences méthodologies[3], les éditeurs ont un but commun : ils cherchent à réduire l'écart entre le texte et le lecteur moderne pour en faciliter l'accès [Breuil, 2019, p. 675].

L'habitude de travailler avec des éditions imprimées amène souvent de jeunes étudiants et étudiantes de lettres à ne pas réaliser que l'édition d'un texte grec,

1. Tentative de reconstitution de l'archétype, soit un texte idéalement originel [Breuil, 2019, p. 300].

2. « Pour des textes en anglo-normand, par exemple, les éditions des historiens n'ont ni accents, ni apostrophes : on lit ainsi *Dangleterre* et *labbe*, au lieu de *d'Angleterre* et de *l'abbé* » [Breuil, 2019, Introduction, p. 5].

3. « L'absence de méthodologie commune a souvent été reprochée à la philologie, mais cette absence est intrinsèque à la discipline ; la philologie consiste plus en un faisceau de règles méthodologiques qu'en une doctrine homogène » [Carles and Glessgen, 2015].

latin ou en ancien français est extrêmement différente de ce qu'on peut lire directement dans le manuscrit, car l'édition en a lissé les difficultés. Afin de faciliter la lecture, la segmentation des mots a été uniformisée [4]. Cette étape n'est pas sans impact sur la réception du texte, notamment concernant les élisions et les tmèses [5]. Comment transcrire *delarbre* quand il n'est pas aisé de déterminer où sont les espaces dans le manuscrit médiéval (illustration 2) ? Faut-il écrire *del arbre* avec une enclise de la préposition et de l'article, fréquente en ancien français, ou *de l'arbre* qui permet une lecture plus aisée ? Ces harmonisations peuvent donner une vision déformée du texte aux novices. Dans les éditions, l'ensemble des abréviations est développé sans nous demander quel est l'impact de ces modifications. Pourtant tout développement est déjà une interprétation. Par exemple, le choix de développer l'abréviation tironienne ⁊ en *et* ou *e* [6] dépend de la date et de l'aire géographique de composition de la copie ou du texte. Pour comprendre le processus d'établissement d'un texte, mais aussi l'exploiter [7], il est important de donner accès aux informations présentes dans les sources avec un minimum d'interprétation [8].

L'édition numérique ouvre aujourd'hui la possibilité de conserver plusieurs strates d'établissement du texte en répertoriant les données préliminaires, ce que l'édition imprimée jusque là ne permettait pas. En effet, l'utilisation de plus en plus répandue dans les communautés scientifiques en Sciences humaines et sociales (SHS) du standard XML TEI [9] [Burnard, 2015] a permis grâce au balisage textuel de consigner des informations de plus en plus nombreuses, mais surtout réexploitables et interrogeables. Concevoir une édition nativement numérique modifie le travail de préparation de l'édition, car contrairement à l'édition traditionnelle où l'imprimé contient l'ensemble des données auxquelles le lectorat aura accès, dans une édition numérique, décrire philologiquement, linguistiquement le texte et en montrer la matérialité ne relève plus d'un même geste. L'objet éditorial peut être démultiplié. Les fichiers XML TEI fonctionnent alors comme une archive de l'ensemble du travail de préparation de l'éditeur qui pourra être ajouté au paratexte de l'édition numérique.

Le travail préliminaire à l'établissement du texte est sauvegardé et les don-

4. Bien souvent, il n'est pas aisé de voir dans les manuscrits si les mots sont séparés ou pas (illustration 1), sans même parler des manuscrits écrits en *scripta continua* (voir le manuscrit carolingien latin 152 de la BnF : `https://gallica.bnf.fr/ark:/12148/btv1b8452765t/f7.item`).

5. Pour désigner une femme enceinte, que retranscrire : *en charga*, comme sur le manuscrit, ou *encharga*, comme l'entrée du dictionnaire ?

6. Forme dialectale de *et* en anglo-normand.

7. Études statistiques d'une *scripta*, du système abréviatif ou des phénomènes d'agglutination.

8. Le retour complet à la source ne sera jamais possible sans une consultation de l'objet physique qu'est un manuscrit.

9. La *Text Encoding Initiative* (TEI) est un consortium qui développe et maintient un ensemble de normes pour la représentation des textes sous forme numérique. Les différentes préconisations liées à ce standard sont décrites dans les *TEIguidelines* qui spécifient les méthodes d'encodage des textes pour qu'ils soient lisibles par la machine et permettent de traduire les principaux principes d'analyse textuelle dans le domaine des sciences humaines, des sciences sociales et de la linguistique.

nées pérennisées, permettant le contrôle de la qualité du travail scientifique. Afin que les données puissent être comprises correctement et être exploitées en dehors du projet, il est également important de documenter la manière dont le corpus a été encodé. Aujourd'hui, une bonne pratique est d'écrire un ODD (*One Document Does it all*) [Rahtz and Burnard, 2013] [10]. Les fichiers XML TEI peuvent ensuite être transformés à condition d'avoir quelques bases en XSLT, en python, en R et en langages web (HTML, CSS, javascript etc.) pour générer des interfaces de lectures, des analyses statistiques de corpus ou encore des graphiques. Les fichiers XML TEI peuvent aussi être réexploités grâce des protocoles plus simples, en utilisant par exemple TEI pusblisher, qui propose une interface graphique pour créer des visualisations de son édition à partir des fichiers XML TEI et éventuellement de l'ODD, ou encore à l'aide de TXM [Heiden et al., 2010] pour faire des études lexicométriques.

Dans un écosystème de science ouverte et partageable [11], l'encodage peut être repris, modifié, augmenté, permettant à d'autres projets d'exploiter les données, ou encore à un nouvel éditeur de partir d'un travail plus détaillé qu'une édition papier qui comporte un grand nombre d'ambiguïtés ou dont les strates de données intermédiaires ont disparu derrière le texte édité. Ainsi, l'édition numérique permet de passer d'un cheminement linéaire qui n'avait pour aboutissement que le texte édité, à un cheminement ouvert, où les différentes strates de l'établissement constituent des données. L'éditeur numérique peut alors se penser à la fois comme un producteur de texte et comme un producteur de données pour d'autres projets.

Le passage à un format numérique, à notre sens, n'amène pas à faire table rase du passé [Mounier, 2010] ni à révolutionner les méthodologies de l'édition qui peuvent déjà être très différentes d'un courant à un autre, d'un champ à un autre [Duval, 2015, pp. 4-9]. Toutefois, cet enrichissement numérique s'accompagne d'une augmentation du temps consacré à l'annotation du texte, tâche jusqu'alors inexistante, et aujourd'hui obligatoire. Il semble alors d'autant plus important de valoriser ces données invisibles afin qu'elles puissent servir non seulement dans le cadre de l'édition et de l'interprétation du texte, mais aussi pour qu'elles aient une vie en dehors de l'édition. En nous basant sur les travaux que nous avons réalisés dans le cadre de notre thèse [Pinche, 2021], nous exposerons ici trois cas de valorisation et d'exploitation des données numériques : l'exploitation des variantes textuelles, des données linguistiques et des données issues de la transcription graphématique [12].

10. Document intégralement écrit en XML qui fonctionne comme un schéma pour non seulement régir l'encodage et assurer l'homogénéité d'un projet, mais aussi produire une documentation sur les choix d'encodage, pour aller plus loin, consulter le chapitre *Getting Started with P5 ODDs* dans les *TEIguidelines*.

11. Quoiqu'il soit encore naissant aujourd'hui

12. Représentation imitative du texte de la source qui ne tient pas compte des variantes de forme des graphèmes [Stutzmann, 2011].

1 Exploitation des données issues des variantes dans la tradition manuscrite

La collation des différentes variantes présentes dans les manuscrits est nécessaire à l'établissement d'une édition critique. Son but est de constituer un *stemma* afin d'organiser la tradition en familles de manuscrits et de hiérarchiser les témoins en fonction de leur proximité avec un original perdu. Cette méthode est née des théories allemandes de la fin du XIXe siècle et notamment des travaux de K. Lachmann [Fornaro, 2011, Trovato and Reeve, 2014]. P. Mass [Maas, 1960] a fait tendre cette approche vers une étude statistique. La philologie italienne, avec G. Pasquali [Pasquali, 1934], G. Contini [Contini, 1986], ou encore C. Segre [Segre, 2015], s'est emparée de ces travaux pour fonder l'école neo-lachmanienne et a approfondi l'étude de la tradition manuscrite en montrant à quel point la transmission du texte dans les manuscrits pouvait être complexe à cause de la mouvance textuelle et en proposant des solutions au scepticisme de J. Bédier [Bédier, 1928b, Bédier, 1928a, Bédier, 1976] quant à la possibilité d'établir des stemmata pour les textes médiévaux [Roelli, 2020, Introduction p. 4].

Le standard TEI dispose de balises spécifiques pour constituer un apparat critique [13]. L'encodage permet à la collation, devenue données, d'être exploitée au-delà de la simple visualisation de l'apparat [14]. L'intégralité de la collation peut être vérifiée, aisément corrigée, voire augmentée si besoin. L'apparat peut être interrogé et parcouru. Plutôt que de faire un relevé « manuel » des variantes pour créer son stemma [15], la tâche peut être automatisée pour partie. Si une typologie des variantes a été établie au préalable dans l'apparat numérique, on peut alors générer automatiquement une base de données pour étudier statistiquement le corpus, à l'aide de langage comme R ou Python, et produire de manière semi-automatique un stemma.

Pour profiter pleinement des nouvelles possibilités d'exploration du corpus qu'offrent les technologies numériques, l'analyse peut être assistée par le package stemmatology [Camps, 2019] pour R développé par F. Caffiero et J. B. Camps [Camps and Cafiero, 2018]. La méthode proposée est basée sur les principes néo-lachmaniens de généalogie textuelle [16]. L'algorithme s'appuie sur les lieux variant communs aux différents témoins et par un jeu de comparaison distingue les leçons ayant un caractère généalogique du bruit généré par les variantes peu significatives. Une fois le tri opéré [17], des relations de parenté sont établies entre les manuscrits. Reléguer une partie du tri et de la compa-

13. Voir les *TEIguidelines* sur l'encodage des apparats critiques.

14. Exemple d'une page d'édition critique issue de la *Vie de saint Martin*, figure 3.

15. Étude qui par ailleurs est toujours extrêmement subjective puisqu'il faut sélectionner un certain nombre de variantes signifiantes qui soit traitable par l'homme.

16. Cette méthode découle elle-même de la méthode Lachmanienne qui s'appuie sur les erreurs communes des témoins d'une tradition pour déterminer leurs liens de parenté entre les différents manuscrits.

17. Le tri s'effectue de manière semi-automatique, dans les cas où une leçon est problématique, l'algorithme permet un choix manuel qui peut être éclairé par la connaissance de la tradition manuscrite.

raison des lieux variant à la machine revêt de nombreux avantages, d'autant plus lorsque l'on travaille avec des textes en langue vernaculaire dont les variations et les remaniements sont foisonnants d'un témoin à un autre. Cette profusion a d'ailleurs bien souvent découragé la critique textuelle moderne face au travail titanesque et parfois peu concluant que représente l'établissement des stemmata de ces œuvres [18]. L'algorithme permet, en parcourant l'ensemble des données collectées [19], de faire ressortir les lieux variants déterminants. Grâce à cette méthode, le processus d'établissement du stemma a pu prendre en compte 709 lieux variants [20] sélectionnés dans la *Vie de saint Martin* [21], certaines leçon, évaluées par l'algorithme comme étant problématiques dans la tradition et très certainement à la racine de divergences entre deux clusters (voir figure 5), nous permettent de voir très clairement se dessiner deux branches dans la tradition (voir exemple 1), mais aussi de faire apparaître des phénomènes de contamination (voir exemple 2). Ainsi un stemma final a été établi (voir figure 4) en accord avec les précédentes recherches sur les légendiers hagiographiques français de Paul Meyer [Meyer, 1906] et les hypothèses de J. J. Thompson [Thompson, 1993], tout en permettant aussi d'approfondir l'analyse et de révéler la position toute particulière du manuscrit F^2 [22], un point de contagion entre les deux branches dans la tradition manuscrite [23].

Exemple 1 : Sélection de leçons qui révèlent une configuration en deux sous-réseaux

— *exemple prendre* (C^1, C^2, C^3, D, E2, F2), *prendre essample de bien* (G1, M1, N1)
— *flamber* (C1, C2, C3, D, E2, F2), *faire grant flambe* (G1, M1, N1)
— *la fermeüre desfermer del huis* (C1, C2, C3, D, E2), *defermer huis* (F2), *la serrure desfremer* (G1, M1, N1).

Exemple 2 : Sélection de leçons qui révèlent des contaminations entre les deux sous-réseaux

— *De diversitez* (C1, C2, C3, D, E2), *d'aversitez* (F2, G1, M1, N1)
— *qi de Poitiers estoient* (C1, C2, C3, D, E2), *de Poitiers* (F2, G1, M1, N1)

Cette méthode présente toutefois quelques limites. Il n'existe pas encore de liste(s) de types de variantes pour faciliter le moissonnage et surtout mettre en

18. « *It has been argued by some modern textual critics that coincident and inevitable that the formation of a stemma is impossible and any attempt to do so is a waste of time* », [Poole, 1974].

19. Ce qui serait extrêmement coûteux en temps et en main-d'œuvre sans une assistance computationnelle. On peut, ici, se référer aux expériences de Dom Quentin et Dom Froger qui ont commencé « à la main » de telles expériences avant d'adapter leur méthode à une pratique automatisée et assistée par l'ordinateur [Quentin, 1926, Froger et al., 1968, Froger, 1970].

20. Le nombre de lieux variants sélectionnés aurait difficilement pu être étudié manuellement dans le cadre d'un travail individuel limité dans les bornes temporelles d'un doctorat

21. Seules les variantes ayant été identifiées dans l'encodage comme étant des variantes sémantiques ont été utilisées.

22. Voir la liste des manuscrits utilisés dans la collation en annexes, tableau 1.

23. Nous remercions J. B. Camps qui nous a accompagnées tout au long de l'exploration du corpus avec *Stemmatology*.

place un encodage homogène d'un projet à un autre [24]. Un tel classement permettrait également d'améliorer des outils comme *stemmatology* qui pourraient alors proposer de pondérer certains types de leçons, évitant ainsi une sélection subjective d'un certain nombre de variantes parmi les plus signifiantes ou parmi celles qui proposent un texte difficile à reproduire, à corriger, ou encore une altération sémantique.

En France, les éditions critiques imprimées ne consignent que les variantes sémantiques [25]. Dans le cas des éditions de textes en langue vernaculaire, on comprend d'autant mieux ces restrictions que la variété des graphies rend impossible l'affichage de l'intégralité des variations dans la limite d'une page papier. Comme le souligne F. Duval au sujet des éditions numériques et de leur rapport à la *new philology* :

> « Pour des raisons économiques, le papier est contraint de se limiter à un seul état textuel (*single text editions*, *one text editions*), qu'il soit ou non reconstruit. Grâce au "multi-fenêtrage" et à l'hypertexte, le numérique s'est imposé comme le médium approprié et indispensable à l'application des "nouvelles" théories textuelles. » [Duval, 2017]

L'apparat numérique permet de dépasser ces limites matérielles et de transmettre une collation plus détaillée avec les variations de nombre, de genre, les variations sur les mots outils, voire même les variations graphiques [26]. Dans la lignée des nouvelles théories philologiques qui accordent une plus grande importance à la copie que ne le faisaient les courants de la fin du XIXe et du XXe siècle, les éditions numériques permettent de donner davantage d'importance à la matérialité du texte et à sa transmission dans la tradition manuscrite. On peut alors intégrer aisément dans la collation tout un éventail de variations pour analyser la langue des autres manuscrits et les liens que les témoins entretiennent entre eux [27].

Dans l'édition des *Seint Confessor*, l'apparat critique (voir l'exemple 6) se concentre essentiellement sur les variations d'ordre sémantique au niveau lexical, sur les additions, les omissions, les inversions, mais s'intéresse aussi aux variations qui touchent les mots outils, même quand l'impact est moindre d'un point de vue sémantique. Une analyse des variations sur les mots outils (voir

24. Des essais de classification des variantes ont été faits par R. Wilhem et pourraient être repris avec profit [Wilhelm, 2015].

25. Par exemple, la collection *Textes littéraires français* de Droz ou encore la collection *Champion classique « séries moyen âge »* d'Honoré Champion.

26. Toutefois, dans le cadre de l'édition d'un texte en ancien français une telle précision rendrait la collation extrêmement chronophage. Dans ce cas une automatisation de la tâche avec une récupération complète du texte grâce à un outil d'HTR (*handwritten text recognition, reconnaissance de l'écriture manuscrite*) et un alignement automatique des variantes semblerait plus pertinent (voir la communication de J. B. Camps, E. Spadini et L. Ing, Collating Medieval Vernacular Texts. Aligning Witnesses, Classifying Variants) .

27. « Grâce à des approches quantitatives et comparatives, des champs d'études s'ouvraient et les questionnaires scientifiques se renouvelaient : la linguistique diachronique ou d'états anciens en a profité, notamment en syntaxe, par l'étude de la ponctuation ou de la segmentation graphique. L'étude des microvariantes est un secteur en développement [Lepage and Milat, 2008] et l'encodage toujours plus poussé des documents, notamment allographétique, promet de nombreuses découvertes » [Duval, 2017].

exemple) entre les deux principaux représentants de chacune des branches du stemma a révélé une modernisation de la langue.

Exemple de remplacement des mots outils entre les témoins C^1 et G^1 :

— Remplacement de *tresqu(e)* par *jusqu(e)* (10 fois)
— Remplacement de *ensamble (o)* par *avec* (21 fois)
— Remplacement de la forme étymologique *er(en)t* par la forme moderne *estoi(en)t* (20 fois)
— Remplacement de l'expression *avoit non* par *estoit apelez* (6 fois)

Enfin, la collation numérique a permis de dépasser les hypothèses de J. J. Thompson qui fut le premier à étudier les *Seint Confessor* et à en comprendre le fonctionnement [Thompson, 1993]. En effet, l'approche quantitative par compte de mots a prouvé que G^1, principal représentant de la deuxième branche, ne donnait pas une version abrégée du recueil, mais une version remaniée. En effet, il supposait que la version de C^1, principal représentant de la première branche, était une version longue du recueil des *Seint Confessor*, tandis que G^1 en proposait une version courte [Thompson, 1993, p. 45-56]. Toutefois, si on compare les deux recueils au niveau des unités textuelles, le nombre de mots utilisés entre les deux versions de la *Vie de saint Martin* révèle pour G^1 une diminution de seulement 2 %. L'analyse plus fine des lieux variant montre que les deux témoins proposent deux versions différentes du texte qui pourraient mériter une édition synoptique, car presque 45 % du texte se trouve modifié entre les deux rédactions. La *Vie de saint Martin* possède un peu moins de 2 300 endroits où les deux manuscrits proposent des leçons divergentes. Environ 27 % des modifications sont constituées par des omissions et 52 % par des variations sémantiques (voir la répartition de types de modification en fonction des sections de la *Vie de saint Martin*, figure 7) [28].

Grâce à l'adoption des technologies numériques, certains projets comme le projet *Hyperdonat* ont expérimenté de nouvelles voies en proposant de parcourir la tradition manuscrite à travers une interface de comparaison des témoins, mais aussi une interface de reconstitution de « virtual witnesses » à partir des témoins existants de la tradition [29]. Le projet d'édition de *Guiron le courtois* mérite également toute notre attention. Le texte numérique est établi branche par branche avec des éditions intermédiaires pour chacune d'entre elles afin d'établir un texte critique qui garde une trace de la surface linguistique pour chaque branche de la tradition. La réunion des différentes éditions permettra à terme de retracer le stemma dans toute sa complexité et petit à petit d'arriver à l'édition critique finale de l'œuvre tout en suivant les innovations au fil de la transmission du texte [Trachsler and Leonardi, 2015].

Dans le cadre de nos travaux, les données de la collation ont pu être exploitées

28. [Pinche, 2021, pp. 120-123].

29. Le projet, encore expérimental et malheureusement inachevé, propose la mise en place d'une matrice d'apparat assez complexe à partir de la collation de toutes les variantes des témoins selon la typologie suivante : *semantic*, *graphic*, *layout*, *structure*. L'ensemble des leçons de chacun des témoins s'appuie sur le texte établi dans les éditions de Paul Wessner, puis les leçons de chacun des témoins sont rassemblées pour former une matrice de variantes qui permet la mise en place de l'édition critique, de l'interface de comparaison des témoins ou encore de créations de « virtual witnesses » [Pinche et al., 2016].

en dehors de l'apparat critique et même dans un autre but que l'établissement du stemma. Elles ont permis de constituer une étude chiffrée des divergences entre deux versions du texte et de consolider le commentaire de l'œuvre. Là où ces informations seraient restées cachées dans les brouillons de l'éditeur, grâce à l'édition numérique, elles ont pu être consignées dans les fichiers XML TEI, archivées et réutilisées.

2 Exploitation des données linguistiques

Le texte établi dans l'édition des *Seint Confessor* ne propose pas de lissage linguistique et dans la mesure du possible, sauf erreur évidente du scribe, livre le texte du manuscrit de base [30] avec le moins de correction possible. Si cette méthode est parfois contestée, notamment par les neo-lachmanniens, car elle ne propose pas de reconstruire le texte le plus proche possible de l'archétype et est parfois ressentie comme un refus d'éditer, comme me constate E. Pierazzo [Pierazzo, 2015], nous avons préféré proposer un texte qui soit le témoignage d'un document historique avec ses irrégularités linguistiques dans la lignée des propos de C. Marchello Nizia :

> « Notre but est de procurer une version du texte sous une forme la plus fidèle possible à la version singulière transmise par le manuscrit de base choisi. C'est à cette condition que nous pouvons accéder à un témoin effectif, à une version précise du roman, telle qu'elle a été lue, écoutée, recopiée sans doute au XIIIe siècle : à ce que l'on pourrait nommer une version "usagée" du texte imparfaite peut-être, mais transmise par un copiste et reçue par des lecteurs-auditeurs. Il s'agit d'un choix bien pesé, et sans doute influencé par le fait que les médiévistes à l'origine du projet étaient des linguistes, et tout spécialement des linguistes diachroniciens qui ne pouvaient concevoir l'accès aux changements linguistiques et à l'évolution des langues autrement qu'à travers l'exploration de témoins indiscutables, c'est-à-dire d'énoncés ayant été réellement performés dans un échange précis [Marchello-Nizia, 2019, p. 59] »

Ce choix a permis une analyse linguistique et statistique d'un corpus « de terrain », ainsi que d'obtenir un relevé exhaustif des phénomènes à étudier. En regard de l'étendue de notre projet numérique, 123 000 tokens [31], la tâche a été automatisée. Un étiquetage automatique des lemmes et de la morphosyntaxe a été mis en place à l'aide d'un annotateur qui fonctionne à partir d'algorithmes de *deep learning* (apprentissage profond) qui ne s'appuient pas sur un dictionnaire ou un ensemble de règles prédéfini en raison des variations propres au français médiéval. L'annotateur *Pie* [Manjavacas et al., 2019] pour l'ancien français fonctionne par apprentissage machine à partir d'un corpus d'entraînement annoté [32]

30. Manuscrit C^1
31. Dans le cadre de notre étude, un token est un élément du texte obtenu suite à la tokenisation qui peut aussi bien être un mot qu'un signe de ponctuation.
32. Principes de *Machine learning* supervisé.

et associe à chaque mot un lemme, une POS (*Part Of Speech*) et une analyse morphosyntaxique [Pinche, 2019]. Dans un deuxième temps, l'annotation a été vérifiée grâce à l'interface de post-correction de Pyrrha [Clérice et al., 2018] disponible en ligne [33]. L'interface a permis de générer pour chaque texte un fichier XML-TEI annoté [34].

L'annotation linguistique a servi de base pour explorer le corpus. Elle a permis, par exemple, d'étudier la quantité de termes touchés par un phénomène dialectal et de voir que certains phénomènes étaient généralisés [35], tandis que d'autres étaient en réalité assez marginaux. Bien que *Li Seint Confessor* soient identifiés comme appartenant à la sphère linguistique des dialectes du Nord-Est, où le son [ọ] est généralement noté <iau> ou <au>, les formes en <iau> n'apparaissent qu'en de très rares endroits, au bénéfice de la forme en <eau>, identifiée comme francienne [36]. Par exemple pour le terme « beau », on trouve 7 occurrences de la graphie « biau(s) », contre 117 de la graphie « beau(s) ». Ainsi, l'étude de la scripta d'un texte nous montre qu'un manuscrit offre toujours à son lecteur un état de la langue complexe issu de strates linguistiques diverses, mais aussi que notre connaissance des corpus médiévaux est encore à approfondir. Une annotation linguistique systématique des textes édités permettrait alors de mettre à jour les grammaires et les encyclopédies linguistiques [Dees et al., 1987] et de poursuivre l'entreprise du Nouveau Corpus d'Amsterdam de Pierre Kunstmann et Achim Stein [Kunstmann and Stein, 2007].

Toutes les données linguistiques du corpus sont maintenant réutilisables, notamment par les linguistes qui voudraient approfondir son exploration. La génération automatisée de ces données amène à imaginer des élargissements. En utilisant les mêmes technologies, l'étude linguistique pourrait être étendue aux autres témoins de la tradition pour comparer les résultats et essayer de repérer les traits récurrents d'une copie à une autre, et donc peut-être ceux propres à l'auteur. Une autre étude à l'échelle d'un manuscrit complet permettrait, quant à elle, de comparer les différents textes et peut-être de faire surgir les traits liés cette fois au scribe, le problème restant dans les deux cas l'acquisition du texte pour étendre l'étude.

L'annotation linguistique a également pu être réutilisée en dehors de ce contexte. Elle a permis de faire quelques prospections lexicométriques, en utilisant l'indice de spécificité de Lafon [Lafon, 1980] et TF-IDF [37] (*term frequency-inverse document frequency*) afin de déceler des thèmes récurrents et spécifiques à notre texte pour faciliter l'analyse le recueil. Grâce aux études lexicométriques, on remarque la prégnance de nouveaux termes dans les trois dernières Vies des

33. `https://dh.chartes.psl.eu/pyrrha`

34. Chaque mot est englobé dans une balise *w*. Le lemme est indiqué grâce à l'attribut *lemma* et l'étiquetage morphologique est quant à lui consigné dans les attributs *pos* et *msd*, suivant les préconisations des TEIguidelines, 17.4.2 *Lightweight Linguistic Annotation*, voir figure 9.

35. Par exemple, la réduction picarde de *-iee* en *-ie* pour les participes féminins.

36. *Cf.* Gossen §12, Pope §501, §1320 xvii et §1322 ix

37. La méthode permet de confronter la fréquence d'un terme dans un texte à sa fréquence dans un corpus plus large, [Sparck Jones, 1972, pp. 11-21].

Seint Confessor tels que *frere*[38], *moine*[39] et *abeie*[40] dans les Vies de saint Benoît et de saint Jérôme où l'accent est mis sur la vie en communauté. Dans la *Vie de saint Alexis*, on voit se dégager l'idée de renoncement au monde à travers le terme *povre*[41]. Ces termes témoignent d'une présence de plus en plus forte dans le recueil d'une idéologie de l'ascèse (voir tableau 2 pour les scores TF-IDF et tableau 3 pour les scores de l'indice de spécificité de Lafon).

Toutefois, le travail de lemmatisation et surtout de correction des données, si l'on veut atteindre une qualité acceptable pour une analyse linguistique sur des graphies aussi peu stables que celles de l'ancien français, est un travail extrêmement long[42]. Il demande beaucoup de précision et peut se révéler extrêmement peu gratifiant[43], même s'il est fondamental pour la connaissance de notre corpus. Dans notre cas, ce travail n'est pas resté enfermé dans les limites de l'édition. Ces efforts ont permis de constituer, pour la lemmatisation de textes en langue d'oïl, un corpus dit « gold »[44]. Ce corpus s'appuie sur deux standards. l'étiquetage morpho-syntaxique et la constitution de ses catégories s'appuient sur le référentiel étendu de *Cattex 2009* [Guillot et al., 2013]. Les lemmes ont été établis à partir du dictionnaire *Tobler-Lommatzsch* [Tobler and Lommatzsch, 1952] et adaptés dans les cas où les entrées n'étaient pas homogènes[45]. Le respect de ces standards a permis aux données de s'intégrer dans un ensemble de corpus afin de créer un modèle général pour la lemmatisation de l'ancien français et d'entraîner le modèle d'annotation *Deucalion* développé à l'École nationale des chartes [Clérice et al., 2020]. Aujourd'hui le modèle atteint un niveau de fiabilité de 95 % pour les lemmes et les POS et de 90 % pour l'annotation morphosyntaxique[46].

Ainsi, la lente accumulation des données linguistiques n'a pas disparu sous le commentaire linguistique, mais au contraire a pu être réutilisée pour le commentaire de l'œuvre. Les données sont aussi sorties de l'édition et ont déjà démontré leur utilité en aidant à créer un modèle de lemmatisation qui pourra servir à automatiser la tâche d'annotation linguistique pour d'autres corpus (voir le schéma de la réutilisation des données issues de la lemmatisation, figure 7).

38. *Vie de saint Benoît* : Score TF-IDF : 0.106 – spécificité de Lafon 90,1. *Vie de saint Jérôme* : score TF-IDF : 0.14 – spécificité de Lafon 15,4.

39. Le terme ne ressort tout particulièrement dans la *Vie de saint Benoît* avec un score TF-IDF de 0,073 et Lafon de 47,4.

40. *Vie de saint Benoît* : Score TF-IDF : 0.069 – spécificité de Lafon 39,2. *Vie de saint Jérôme* : score TF-IDF : 0.04 – spécificité de Lafon 3,6.

41. Score TF-IDF : 0.079 – spécificité de Lafon 8,5.

42. Sur un corpus identique au notre (123000 tokens), à un rythme de 200 mots par heure, cela représente un peu moins de 90 jours de 7 heures de travail uniquement appliqués à cette tâche

43. En cela, des outils de post-correction comme Pyrrha sont d'une grande aide.

44. « *Gold standard is a dataset which has been annotated (either manually or automatically) and then manually corrected* », `https://port.sas.ac.uk/mod/book/view.php?id=612&chapterid=426`.

45. Voir la documentation à l'adresse suivante : `https://github.com/Jean-Baptiste-Camps/Geste/wiki/5BRÉF%5D-lemmes:-lemmes-retenus-de-Tobler-Lommatzsch`

46. Seuls 30 % des données du modèle possèdent cette information qui, par ailleurs, est plus complexe à traiter automatiquement en raison de sa forte disparité.

3 Exploitation des données issues de la transcription graphématique

Afin de respecter le texte comme document, les fichiers XML TEI sources qui ont servi à l'établissement du texte conservent en leur sein toutes les données de mise en page, ainsi qu'une transcription graphématique du manuscrit de base. Consigner toutes ces informations a permis de proposer aux lecteurs plusieurs vues du texte (voir figures 12 et 13), ce qui permet de vérifier les informations d'établissement du texte et de faciliter le retour à la source pour tout lecteur non expert en paléographie. La représentation imitative du texte (voir figure 12) renferme des indications de mise en page qui peuvent aider à comprendre l'origine des erreurs de copie [47]. Les abréviations présentes dans le document source sont également signalées, ainsi que la ponctuation originale (voir figure 11). Ces informations ne sont pas uniquement destinées à permettre un affichage imitatif du texte du manuscrit de base, mais pourront être réexploitées par des linguistes pour des études sur des phénomènes restreints, à l'instar de l'étude sur la ponctuation menée par A. Lavrentiev [Lavrentiev, 2016].

Cette méthode a permis d'établir un rapport précis des modifications opérées. Par exemple, le corpus des *Seint Confessor* comporte 6300 abréviations sur un corpus de 67 folios à deux colonnes. L'utilisation de la note tironienne ⁊ représente 75% des abréviations et 92 % des abréviations sont répartis entre sept signes différents. Outre le fait que ces informations permettent d'accompagner l'édition d'annexes précises sur la manière dont les abréviations ont été développées dans le corpus (voir tableau 4), ces statistiques montrent que le texte en langue vernaculaire peut être aisément lu sans une connaissance savante du système abréviatif. Ainsi les fichiers XML TEI, qui peuvent être enrichis de liens vers des reproductions de manuscrits, possèdent non seulement des vertus pédagogiques, mais permettent aussi la mise au point d'une transcription précise dont la production s'avère d'une grande utilité pour la connaissance du texte.

Ces données peuvent également être réutilisées dans un écosystème numérique plus large (voir figure 14), notamment pour l'entraînement d'algorithmes de reconnaissance automatique de texte (HTR) [48]. L'encodage en XML-TEI des fichiers de notre projet en consignant les abréviations du manuscrit, les sauts de ligne, la mise en page des colonnes et des folios a permis de fournir un set de données pour entraîner en 2019 un modèle d'HTR [49] (voir figure 14) afin de

47. Par exemple, on lit dans C^1 au folio 130d, dans la *Vie de saint Gilles* : « *Qant li rois vit qe la chose (estoit) estoit veraie* ». L'erreur par dittographie du scribe peut d'expliquer par un saut de ligne entre les deux *estoit*.

48. Entraîner soi-même un modèle d'HTR demande d'être capable de lancer quelques commandes en Python. Aujourd'hui des services permettent de le faire à l'aide d'interface graphique, comme eScriptorium [Kiessling et al., 2019] qui donne librement accès aux modèles entraînés ou encore Transkribus qui en revanche est propriétaire des modèles.

49. Malheureusement notre travail de thèse n'a pas pu bénéficier d'une aide à la transcription via l'utilisation de l'HTR dès son début. Le travail ayant été entamé en 2015, les modèles et les données librement disponibles pour les manuscrits médiévaux (nous avions surtout testé Kraken pour pouvoir entraîner nos propres modèles) ne permettaient pas encore d'obtenir des données d'assez bonne qualité pour être intégrées dans la chaîne de production du texte.

produire une transcription automatisée de l'ensemble du manuscrit fr. 412 de la BnF, et d'étudier la composition du légendier complet grâce à un protocole allant de l'acquisition du texte jusqu'à l'analyse stylométrique [Pinche et al., 2019]. Grâce à ce protocole, nous avons analysé la composition du manuscrit et retrouvé des sous-collections hagiographiques cohérentes avec les hypothèses de Paul Meyer selon lesquelles les légendiers sont le fruit d'assemblage de compilations successives [Meyer, 1906], laissant ainsi penser que certaines vies pouvaient être du même auteur et fonctionner, à l'instar des *Seint Confessor*, comme des ensembles cohérents.

Aujourd'hui, l'apprentissage machine a fait beaucoup de progrès et des volontés de partage de données pour l'HTR voient le jour comme le projet *HTR-united*. Il est important que les éditeurs numériques trouvent leur place dans un environnement scientifique où il est fort probable que l'acquisition automatique de texte prendra de plus en plus de place. Les transcriptions des *Seint Confessor* ont servi de base à la constitution d'un ensemble de données d'entraînement, disponible sur le dépôt Github *Cremma-medieval*, en vue de créer un modèle HTR plus général pour les manuscrits littéraires des XIIe et XIIIe siècles [Pinche and Clérice, 2021] et ont permis de constituer un modèle d'HTR (Bicerin 1.0.1) avec des scores de 95,49 % d'*accuracy* sur un jeu de données composé de huit manuscrits différents [50]. L'apparition de ces technologies soulèvent de nouveaux enjeux et questionne la place de l'éditeur dans la chaîne d'acquisition textuelle qui pourrait être de plus en plus automatisée de l'acquisition à la visualisation du texte [Chagué and Chiffoleau, 2021]. Des réflexions ont également été menées par J. B. Camps, L. Ing et E. Spadini lors de la conférence *Digital Humanities* en 2019 sur une chaîne de production automatisée allant de l'acquisition automatique du texte jusqu'à de l'alignement des témoins et la classification de leurs lieux variants [Camps et al., 2019].

Il est donc, à notre avis, important que les compétences d'éditeur soient mises au service de la confection de transcriptions pour les données d'entraînement de l'HTR et que, pour en assurer la qualité, commencent à émerger parmi eux des groupes de réflexion sur les nouvelles problématiques de transcription afin de créer les données textuelles les plus qualitatives possibles, mais aussi les plus adaptées à l'apprentissage machine. De nouveaux enjeux apparaissent, comme par exemple, l'homogénéisation et le choix des caractères spéciaux [51]. Ces questionnements devraient aboutir pour une grande partie à une explicitation et une homogénéisation des méthodologies déjà décrites par certains guides d'éditions imprimées [52]. L'établissement de transcriptions pour des corpus numériques se révèle être un véritable champ scientifique en chantier, qui loin d'être accessoire, est très complémentaire d'études de qualité, car, de fait, la composition de corpus numériques qui puissent être compatibles entre eux nous force à expliciter la question de la segmentation des mots, des ajouts de signes diacritiques, du

50. Ce modèle sert déjà à d'autres projets pour accélérer le processus d'acquisition textuelle, nous pouvons citer ici le projet genevois en cours : *Canoniser les Sept Sages*.

51. Voir à ce sujet le projet *MUFI : The Medieval Unicode Font Initiative*.

52. [Bourgain and Vielliard, 2001], [Bourgain and Vielliard, 2018], [Vielliard and Guyot-jeannin, 2001], [Lepage, 2001].

lien entre le texte et l'objet [53].

Conclusion

L'édition numérique permet de nous ramener à la source et de proposer des descriptions toujours plus complexes et fournies. L'établissement d'un texte numérique, contrairement à certaines idées reçues, permet rarement de gagner du temps [54], mais permet de faire vivre des données patiemment collectées en dehors de l'édition, là elles demeurent souvent invisibles dans les éditions imprimées.

Toutefois, n'oublions pas les avertissements d'E. Pierazzo et P. Robinson, l'édition numérique ne doit pas se tourner uniquement du côté de l'édition documentaire et d'une description sans fin de la source, ce qui pourrait être perçu comme un refus d'éditer et amènerait à proposer des projets impossibles à terminer [Pierazzo (eds), 2016, Robinson, 2003, Robinson, 2013]. Il est important de continuer à proposer un texte édité à diffuser à une communauté de lecteurs, chercheurs comme non-spécialistes. L'édition numérique par sa modularité nous offre la possibilité de concilier en son sein une approche du texte comme document et comme œuvre et les éditeurs doivent se saisir de cette opportunité [55].

La production numérique pousse sans cesse le chercheur en dehors de son texte. L'édition numérique le projette vers un au-delà du but visé, elle devient un simple point de départ de recherches qui la dépasse (voir figure 15). L'édition n'est plus qu'une des productions possibles du travail sur le texte. Les corpus numériques nous amènent à passer d'une pratique individuelle qui a pour but de produire un objet fini, qui n'a besoin que d'être compris par quelqu'un d'autre, vers la production de données « en réseau » qui pourront à terme rejoindre, en-dehors de l'édition, des données produites par d'autres projets, ou être réexploitées dans un autre cadre que l'objectif pour lequel elles ont été crées. Ainsi l'éditeur numérique, à notre avis, a pour rôle aujourd'hui d'ouvrir le texte édité à d'autres usages, et doit en cela peut-être modifier ses habitudes pour passer de la production d'un texte clos sur lui-même à la modélisation de données [56] à partager.

53. Pour aller plus loin, voir les travaux de mise en place d'une ontologie pour la description des documents du projet SegmOnto [Gabay et al., 2021].

54. Cependant les améliorations des technologies de transcription et d'annotation automatiques nous permettent d'espérer que le temps consacré à l'acquisition textuelle et à l'annotation linguistique devrait se voir réduit.

55. « *One cannot know the work without the documents – equally, one cannot understand the documents without a comprehension of the work they instance. From this, a principle appears : a scholarly edition must, so far as it can, illuminate both aspects of the text, both text-as-work and text-as-document. Traditional print editions have focused more on the first. An evident advantage of digital editions is that they might redress this balance, by including much richer materials for the study of text-as-document than can be achieved in the print medium* » [Robinson, 2013, p. 123].

56. Les données ne sont pas un acquis naturel, elles sont une construction scientifique dont la qualité est primordiale pour assurer celles des analyses qui les utiliseront, « *Data are capta, taken not given, constructed as an interpretation of the phenomenal world, not inherent in it* » [Drucker, 2011].

La pratique des outils numériques amène à devoir tout expliciter et à interroger ses propres pratiques, car elle demande une grande rigueur et surtout de proposer une méthode reproductible. Il devient donc urgent d'essayer d'homogénéiser la constitution des corpus, la modélisation des données et d'intégrer dans nos pratiques des standards existants. Cependant, cette démarche est encore difficile en l'absence de préconisations ou d'un guide semblables à ceux de l'édition traditionnelle. Ceux-ci seraient, certes, d'autant plus complexes à mettre en place qu'ils s'adressent à un champ où les technologies évoluent rapidement.

Enfin, la valorisation des données numériques de l'édition n'est pas aisée, car il faut les rendre accessibles pour qu'elles trouvent un lectorat capable de les réexploiter. Toutefois, on peut noter l'émergence de quelques habitudes de mise à disposition avec des dépôts sur Github, la création de DOI (*Digital Object Identifier*), notamment via Zenodo, ou encore la mise en place de dépôts par des institutions comme HumaNum avec GitLab [57] et Nakala [58]. Il est également primordial de prendre conscience de l'importance des standards pour assurer la pérennité du travail accompli et la citabilité de son édition et de ses données connexes [59] pour en assurer la réutilisation selon les critères définis par M. Dacos et P. Mounier [Dacos and Mounier, 2010].

Références

[Almas et al., 2021] Almas, B., Clérice, T., Cayless, H., Jolivet, V., Liuzzo, P. M., Romanello, M., Robie, J., and Scott, I. W. (2021). Distributed Text Services (DTS) : a Community-built API to Publish and Consume Text Collections as Linked Data.

[Bourgain and Vielliard, 2001] Bourgain, P. and Vielliard, F. (2001). *Conseils pour l'édition des textes médiévaux. Fascicule II, Actes et documents d'archives*. Comité des travaux historiques et scientifiques : École nationale des chartes, Paris, France.

[Bourgain and Vielliard, 2018] Bourgain, P. and Vielliard, F. (2018). *Conseils pour l'édition des textes médiévaux. Fascicule III, Textes littéraires*. Number 32 in Orientations et méthodes. Comité des travaux historiques et scientifiques : École nationale des chartes, Paris, France.

[Breuil, 2019] Breuil, E. (2019). *Méthodes et pratiques de l'édition critique des textes et documents modernes*. Number 27 in Bibliothèque de littérature du xxe siècle. Classiques Garnier, Paris.

[Burnard, 2015] Burnard, L. (2015). *Qu'est-ce que la Text Encoding Initiative ?* OpenEdition Press, Marseille.

[Bédier, 1928a] Bédier, J. (1928a). La tradition manuscrite du Lai de l'Ombre. Réflexions sur l'art d'éditer les anciens textes (deuxième article). *Romania*, 54(215) :321–356.

57. Plateforme de partage de données.
58. Service de publication, partage et valorisation des données scientifiques.
59. Voir les principes de citation de DTS (*Distributed Text Services*), [Almas et al., 2021]

[Bédier, 1928b] Bédier, J. (1928b). La tradition manuscrite du Lai de l'Ombre. Réflexions sur l'art d'éditer les anciens textes (premier article). *Romania*, 54(214) :161–196.

[Bédier, 1976] Bédier, J. (1976). La philologie médiévale et la critique textuelle. In *Actes du XIIIe Congrès international de linguistique et philologie romanes, tenu à l'Université Laval, Québec, Canada, du 29 août au 5 septembre 1971*. Les Presses de l'Université Laval, Québec, Canada.

[Camps, 2019] Camps, J.-B. (2019). Jean-Baptiste-Camps/stemmatology. original-date : 2014-07-17T12 :34 :16Z.

[Camps and Cafiero, 2018] Camps, J.-B. and Cafiero, F. (2018). Stemmatology : an R package for the computer-assisted analysis of textual traditions. In Frank, A. U., Ivanovic, C., Mambrini, F., and Sporleder, C., editors, *Corpus-Based Research in the Humanities CRH-2*, Proceedings of the Second Workshop on Corpus-Based Research in the Humanities CRH-2, 25-26 January 2018, Vienna, Austria, pages 65–74, Vienna, Austria.

[Camps et al., 2019] Camps, J.-B., Ing, L., and Spadini, E. (2019). Collating Medieval Vernacular Texts. Aligning Witnesses, Classifying Variants. In *DH2019 Digital Humanities Conference 2019*, Utrecht, Netherlands.

[Carles and Glessgen, 2015] Carles, H. and Glessgen, M. (2015). 4. La philologie linguistique et éditoriale. In *Manuel de linguistique française*, pages 108–130. De Gruyter.

[Cerquiglini, 1989] Cerquiglini, B. (1989). *Éloge de la variante : histoire critique de la philologie*. Éd. du Seuil, Paris, France.

[Chagué and Chiffoleau, 2021] Chagué, A. and Chiffoleau, F. (2021). An accessible and transparent pipeline for publishing historical egodocuments.

[Clérice et al., 2020] Clérice, T., Camps, J.-B., Pinche, A., Ing, L., Duval, F., and Kanaoka, N. (2020). deucalion-model-af : 0.3.0.

[Clérice et al., 2018] Clérice, T., Pilla, J., and Camps, J.-B. (2018). hipster-philology/pyrrha : 1.0.1.

[Contini, 1986] Contini, G. (1986). *Breviario di ecdotica*. Einaudi, Milano, Italie.

[Dacos and Mounier, 2010] Dacos, M. and Mounier, P. (2010). *L'édition électronique*. La Découverte, Paris.

[Dees et al., 1987] Dees, A., Dekker, M., Huber, O., and Van Reenen-Stein, K. (1987). *Atlas des formes linguistiques des textes littéraires de l'ancien français*. De Gruyter, Berlin, Boston, reprint 2014 edition.

[Drucker, 2011] Drucker, J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*, 005(1).

[Duval, 2015] Duval, F. (2015). *Les mots de l'édition de textes*. École nationale des chartes, DL 2015, Paris, France.

[Duval, 2017] Duval, F. (2017). Pour des éditions numériques critiques. L'exemple des textes français. In *Le texte à l'épreuve du numérique*, volume 73

of *Médiévales*, pages 13–30. Presses universitaires de Vincennes, Université Paris VIII, Saint-Denis, France, médiévales edition.

[Fornaro, 2011] Fornaro, S. (2011). Karl Lachmann et sa méthode. *Revue germanique internationale*, (14) :125–138.

[Froger, 1970] Froger, D. J. (1970). La critique des textes et l'ordinateur. *Vigiliae Christianae*, 24(3) :210–217. Publisher : Brill.

[Froger et al., 1968] Froger, J., Marichal, R. P., and Faure, R. P. (1968). *La critique des textes et son automatisation*. Dunod, Paris, France. Type : Mémoire.

[Gabay et al., 2021] Gabay, S., Camps, J.-B., Pinche, A., and Jahan, C. (2021). SegmOnto : common vocabulary and practices for analysing the layout of manuscripts (and more). In *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*, Lausanne, Switzerland.

[Guillot et al., 2013] Guillot, C., Prévost, S., and Lavrentiev, A. (2013). Manuel de référence du jeu Cattex09.

[Heiden et al., 2010] Heiden, S., Magué, J.-P., and Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, volume 2, page 1021. Edizioni Universitarie di Lettere Economia Diritto. Issue : 3.

[Kiessling et al., 2019] Kiessling, B., Tissot, R., Stokes, P., and Ezra, D. S. B. (2019). eScriptorium : An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19.

[Kunstmann and Stein, 2007] Kunstmann, P. and Stein, A. (2007). *Le nouveau corpus d'Amsterdam : actes de l'atelier de Lauterbad, 23-26 février 2006*. F. Steiner, Stuttgart, Allemagne. ISSN : 0341-0811.

[Lafon, 1980] Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1) :127–165. Publisher : Persée - Portail des revues scientifiques en SHS.

[Lavrentiev, 2016] Lavrentiev, A. (2016). Ponctuation française du Moyen Âge au XVIe siècle : théories et pratiques. In *La ponctuation à l'aube du XXie siècle. Perspectives his- toriques et usages contemporain*, pages 39–62.

[Lepage, 2001] Lepage, Y. G. (2001). *Guide de l'édition de textes en ancien français*. H. Champion, Paris.

[Lepage and Milat, 2008] Lepage, Y. G. and Milat, C. (2008). *Por s'onor croistre : mélanges de langue et de littérature médiévales offerts à Pierre Kunstmann*. Les Éditions David, Ottawa (Ont.), Canada. ISSN : 1709-8483.

[Maas, 1960] Maas, P. (1960). *Textkritik*. B.G. Teubner, Leipzig, Allemagne.

[Manjavacas et al., 2019] Manjavacas, E., Clérice, T., and Kestemont, M. (2019). pie v0.2.3.

[Marchello-Nizia, 2019] Marchello-Nizia, C. (2019). Édition électronique et introduction linguistique. In *Les introductions linguistiques aux éditions de textes*, pages 55–67. Classiques Garnier, Paris. ISSN : 2257-4700.

[Meyer, 1906] Meyer, P. (1906). Légendes hagiographiques en français. In *Histoire littéraire de la France*, volume 33, pages 328–458. Imprimerie nationale, Paris.

[Mounier, 2010] Mounier, P. (2010). Manifeste des Digital Humanities. *Journal des anthropologues. Association française des anthropologues*, (122-123) :447–452. Number : 122-123 Publisher : Association française de anthropologues.

[Pasquali, 1934] Pasquali, G. (1934). *Storia della tradizione e critica del testo.* Felice Le Monnier, Firenze, Italie.

[Pierazzo, 2015] Pierazzo, E. (2015). *Digital scholarly editing : theories, models and methods.* Ashgate, Farnham Burlington (Vt.).

[Pierazzo (eds), 2016] Pierazzo (eds), M. J. D. a. E. (2016). *Digital Scholarly Editing : Theories and Practices.* Open Book Publishers.

[Pinche, 2019] Pinche, A. (2019). Annoter facilement un corpus complexe. In *Actes des Rencontres lyonnaises des jeunes chercheurs en linguistique historique*, pages 48–58. Association Diachronies Contemporaines, Lyon.

[Pinche, 2021] Pinche, A. (2021). *Edition nativement numérique du recueil hagiographique "Li Seint Confessor" de Wauchier de Denain d'après le manuscrit fr. 412 de la Bibliothèque nationale de France.* Thèse de doctorat, Université Lyon3, Lyon, France.

[Pinche et al., 2016] Pinche, A., Bureau, B., and Nicolas, C. (2016). Hyperdonat, digital edition project.

[Pinche et al., 2019] Pinche, A., Camps, J.-B., and Clérice, T. (2019). Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis. In *Digital Humanities Conference 2019 - DH2019*, Utrecht, Netherlands. ADHO and Utrecht University.

[Pinche and Clérice, 2021] Pinche, A. and Clérice, T. (2021). HTR-United/cremma-medieval : 1.0.1 Bicerin (DOI).

[Poole, 1974] Poole, E. (1974). The Computer in Determining Stemmatic Relationships. *Computers and the Humanities*, 8(4) :207–216. Publisher : Springer.

[Quentin, 1926] Quentin, H. (1926). *Essais de critique textuelle (ecdotique).* Librairie Auguste Picard, Paris, France.

[Rahtz and Burnard, 2013] Rahtz, S. and Burnard, L. (2013). Reviewing the TEI ODD System. *ACM*, pages 193–196.

[Robinson, 2003] Robinson, P. (2003). Where We Are with Electronic Scholarly Editions, and Where We Want to Be. *Jahrbuch für Computerphilologie*, 5(5) :126–146.

[Robinson, 2013] Robinson, P. (2013). Towards a Theory of Digital Editions. In *The Journal of the European Society for Textual Scholarship*, number 10 in Variants, pages 105–131. Brill Rodopi, brill edition. Pages : 105-131 Section : The Journal of the European Society for Textual Scholarship.

[Roelli, 2020] Roelli, P. (2020). *Handbook of Stemmatology : History, Methodology, Digital Approaches*. De Gruyter, Berlin, Allemagne. ISSN : 2698-1998.

[Segre, 2015] Segre, C. (2015). Lachmann et Bédier. La guerre est finie. In Buchi, ., Chauveau, J.-P., Greub, Y., and Pierrel, J.-M., editors, *Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Allocutions de bienvenue, conférences plénières, tables rondes, conférences grand public*, Nancy. ATILF.

[Sparck Jones, 1972] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1) :11–21. Publisher : MCB UP Ltd.

[Stutzmann, 2011] Stutzmann, D. (2011). Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? page 34.

[Thompson, 1993] Thompson, J. J. (1993). *From the translator's worktable to the predictor's lectern : The work of a thirteenth-century author, Wauchier de Denain*. PhD in Candidacy for the Degree of Doctor of Philosophy, Yale University, Yale.

[Tobler and Lommatzsch, 1952] Tobler, A. and Lommatzsch, E. (1952). *Altfranzösisches Wörterbuch*. E. Steiner, Wiesbaden.

[Trachsler and Leonardi, 2015] Trachsler, R. and Leonardi, L. (2015). L'édition critique des romans en prose : le cas de Guiron le Courtois. In Trotter, D., editor, *Manuel de la philologie de l'édition*, pages 44–80. De Gruyter„ Berlin/Boston, Allemagne, Etats-Unis d'Amérique.

[Trotter, 2015] Trotter, D., editor (2015). *Manuel de la philologie de l'édition*. De Gruyter„ Berlin/Boston, Allemagne, Etats-Unis d'Amérique.

[Trovato and Reeve, 2014] Trovato, P. and Reeve, M. D. P. (2014). *Everything you always wanted to know about Lachmann's method : a non-standard handbook of genealogical textual criticism in the age of post-structuralism, cladistics, and copy-text*. Libreriauniversitaria.it edizioni, Limena, Italie. ISSN : 2464-8647.

[Vielliard and Guyotjeannin, 2001] Vielliard, F. and Guyotjeannin, O. (2001). *Conseils pour l'édition des textes médiévaux. Fascicule I, Conseils généraux*. Comité des travaux historiques et scientifiques : École nationale des chartes, Paris.

[Wilhelm, 2015] Wilhelm, R. (2015). L'édition de texte – entreprise à la fois linguistique et littéraire. In Trotter, D., editor, *Manuel de la philologie de l'édition*, pages 131–151. De Gruyter„ Berlin/Boston, Allemagne, Etats-Unis d'Amérique.

[Zumthor, 1972] Zumthor, P. (1972). *Essai de poétique médiévale*. Éditions du Seuil, Paris, France.

4 figures

— title : Exemple de texte issu de la reproduction Gallica du manuscrit fr. 412 de la Bnf, fol. 103r keywords : manuscrit, segmentation type : image lang : fr link : images/01SegmentationMot.png date : 2021 source : Source gallica.bnf.fr / BnF priority : lowpriority —

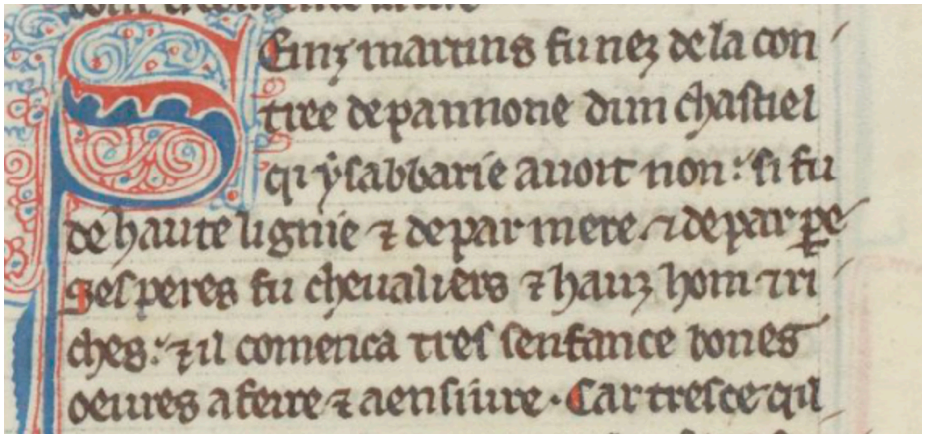


Figure 1 – Exemple de texte issu de la reproduction Gallica du manuscrit fr. 412 de la Bnf, fol. 103r, `https://gallica.bnf.fr/ark:/12148/btv1b84259980/f215.item`

— title : Exemple de texte issu de la reproduction Gallica du manuscrit fr. 412 de la Bnf, fol. 106v keywords : manuscrit, segmentation type : image lang : fr link : images/02SegmentationMot.png date : 2021 source : Source gallica.bnf.fr / BnF priority : lowpriority —

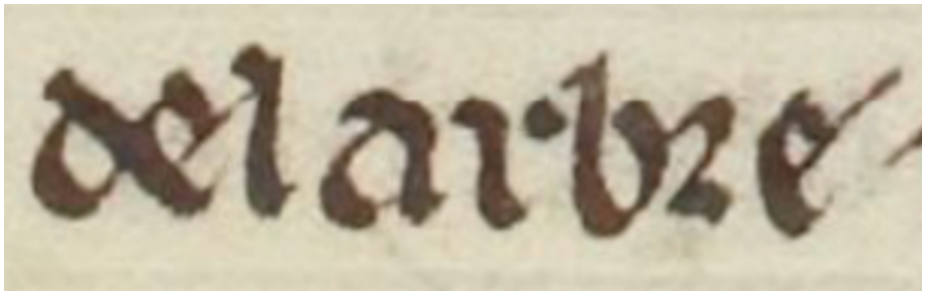


Figure 2 – Exemple de texte issu de la reproduction Gallica du manuscrit fr. 412 de la Bnf, fol. 106v, `https://gallica.bnf.fr/ark:/12148/btv1b84259980/f222.item`

— title : Exemple d'une page d'édition avec apparat critique issu de la Vie de saint Martin keywords : édition, apparat critique type : image lang : fr

Table 1 – Liste des manuscrits utilisés pour la collation de la *Vie de saint Martin*

C^1	PARIS, Bibliothèque nationale de France, Manuscrits, fr. 412, 1285
C^2	PARIS, Bibliothèque nationale de France, Manuscrits, fr. 411, 14e siècle
C^3	LONDON, British Library, Royal 20.D.VI, milieu du 13e siècle
D	PARIS, Bibliothèque nationale de France, Manuscrits, fr. 17 229, 13e siècle
E^2	GENÈVE, Bibliothèque de Genève, Comites Latentes 102, début du 14e siècle
F^2	PARIS, Bibliothèque nationale de France, Manuscrits, fr. 23 117, 13e siècle
G^1	Bruxelles, Bibliothèque royale, 9225, 1re moitié du 14e siècle
M^1	PARIS, Bibliothèque nationale de France, Manuscrits, fr. 23 112, 13e siècle
N^1	PARIS, Bibliothèque nationale de France, Manuscrits, fr. 422, fin du 13e siècle

link : images/03ExempleApparatPapier.png date : 2021 source : auteur priority : lowpriority —

— title : Stemma établi à partir de la collation des variantes de la Vie de saint Martin keywords : variantes, stemma type : image lang : fr link : images/04ExempleApparatPapier.png date : 2021 source : auteur priority : lowpriority —

— title : Exemple de clusters obtenus grâce à Stemmatology keywords : variantes, cluster type : image lang : fr link : images/05ClustersTradition.png date : 2021 source : auteur priority : lowpriority —

— title : Exemple d'encodage de l'apparat critique issu de l'édition numérique de la Vie de saint Martin keywords : XML TEI, apparat type : image lang : fr link : images/06ApparatCode.png date : 2021 source : auteur priority : lowpriority —

— title : Étude de la répartition des types de variantes à partir de la collation de la Vie de saint Martin keywords : collation, variantes type : image lang : fr link : images/07TypesModifParPassage.png date : 2021 source : auteur priority : lowpriority —

— title : Schéma de la réutilisation des données issues de la collation dans l'édition numérique keywords : données, collation type : image lang : fr link : images/08Collation.png date : 2021 source : auteur priority : lowpriority —

1. De seint Martin **m**out doit on doucement et volentiers le bien oïr et entendre, car par le bien savoir et retenir [fol.103b] puet l'en sovent a bien venir. Qui bien ne seit ne bien n'entent de bien faire n'a nul talent. Mes del bien nest sovent li biens, del mal li maus si com dist l'Escriture. Por ce se doit l'en au bien avoier et le bien feire, si com li seint home firent ça en arriere de cui nos trovons les oevres et les vies [es] Escriptures. Et bien sacent tuit cil qi vivent qe ja n'auront tant de bien fet en totes lor vies qe, qant la mort dont nule rien n'eschape les poindera au cuer, q'il ne cuident petit avoir fait. Dex ! Qe feront dont cil qui riche sont et aise de l'avoir de cest siecle, ne en eus n'ont douçor ne humilité ne misericorde, ainz sont plein d'angoisse et de traïsson et de felonie et de si grant avarice qe com plus ont richesces et avoirs, plus en desirrent a avoir ? Ce fet li deables qi en tel maniere les a laciez et pris q'il les enmeine en infer le grant chemin plenier. De ce se gardent li seint home qi par dolereuses peinnes et par griez tormenz et par veilles et par geunes et par toutes bones oevres firent tant q'il vindrent a vie parmenable et a la corone de gloire. A ce regarderent li seint confessors et messires seinz Martins dont ci comence la vie.

2. **S**einz Martins fu nez de la contree de Pannone, d'un chastiel qi Ysabbarie

1 De seint Martin] Ci commence la vie de monseigneur seint Martin *C²*, Ci comence la vie seint Martin *C³*, Ci commence la vie monseignor saint Martin le bon ami nostre Seignor *D*, Ci conmance la vie monseigneur saint Martin qui fu arcevesques de Tours et puis emprés ses miracles ensuianz *E²*, *om. F² G¹ N¹*, C'est la vie de saint Martin vesque *M¹* || **1–4** mout doit on [...] mal li maus] Chascuns doit volentiers oïr le mouteplïement et le bien entendre qui puet venir de bones paroles, car par le bien savoir et retenir ne puet on s'amender non. De bien nait bien et de mal li maux *G¹* || **1–2** mout doit on [...] oïr et entendre] Cascuns crestiens doit bien oïr et entendre volontier le bien *M¹*, Cascuns crestiens doit bien entendre et volentiers oïr le bien *N¹* || **1** et volentiers] *om. F²* || **2** savoir] savoir oïr *E²* || **3** de bien faire n'a nul talent] *om. F²* || **3** nest] vient *N¹* || **3** sovent] adiés *N¹* || **4** del] des *E²* || **4** mal] *om. E²* || **4** se] *om. F²* || **4–5** au bien avoier et] *om. F²* || **4** au bien] *om. E²* || **5–6** les oevres et] *om. F²* || **6** es Escriptures *C² C³ D F²*] escriptures *C¹*, *om. E²*, et les escriptures *G¹*, escrites *M¹*, en escriture *N¹* || **7** en totes lor vies] *om. F²* || **7** totes] *om. M¹* || **7** vies] vivant *M¹ N¹* || **7** la mort] la mort vient *E²* || **7** rien] riens vivant *D*, *om. F²* || **7** poindera] prendra *D E² G¹ M¹*, prend *F²* || **8** ne] n'en *F² M¹* || **8** petit avoir fait] petit de bien fait *N¹* || **8–9** et aise de l'avoir de cest siecle] *om. F²* || **9** aise] assasé *M¹* || **9** ont] a *M¹*, n'en n'ont *N¹* || **9** douçor] *om. F²* || **9** humilité] debonnereté ne humilité *E²* || **10** d'angoisse et] *om. F²* || **10** si] *om. F²* || **10** com] *om. M¹* || **11** richesces et avoirs] *om. F²*, d'avoir et *G¹*, avoir *N¹* || **11** en] *om. F²* || **11** desirrent] covoïtent *F² N¹* || **11** a] *om. F² G¹* || **11** avoir] *om. F²* || **11–12** Ce fet li [...] grant chemin plenier] *om. F²* || **11** deables] anemi *M¹ N¹* || **11–12** en tel maniere] ci *E²* || **12** a] en a *E²* || **12** et pris] et pris en tele maniere *E²*, *om. G¹*, et loiés *M¹ N¹* || **12** q'il les enmeine] qui les enmeine *C² E² G¹ M¹ N¹* || **12** le grant chemin plenier] le grant chemin plenier et la grant voie batue *G¹ M¹ N¹* || **13** gardent] garderent *D*, gardoient *G¹*, garderent bien *M¹ N¹* || **13–14** et par griez tormenz et par veilles et par geunes] *om. F²*, et par martire *G¹* || **13** tormenz] martires *M¹ N¹* || **14** toutes] *om. E² F²* || **15** la] *om. G¹ M¹ N¹* || **15** A ce regarderent] a ce pristrent garde *G¹ M¹ N¹* || **15** regarderent] regarda *F²* || **15** li seint confessors et] *om. F²* || **15** seint] *om. G¹* || **15** confessors] home confes *N¹* || **15** et] meesmement *M¹* || **16** dont ci comence la vie] dont je vous dirai la vie *G¹*, dont je vos conterai ichi la vie *M¹*, dont jou commencerai la vie *N¹* || **17** de la contree] *om. F²*, en la cité *G¹* || **17** Pannone] Paunone *D N¹*, Pauone *F²*, Pantione *E²*, Pannonie *G¹* || **17** Ysabbarie] Ysapbarie *D*, Ysapharie *E² F² G¹ M¹ N¹*

FIGURE 3 – Exemple d'une page d'édition avec apparat critique issu de la *Vie de saint Martin*, [Pinche, 2021]

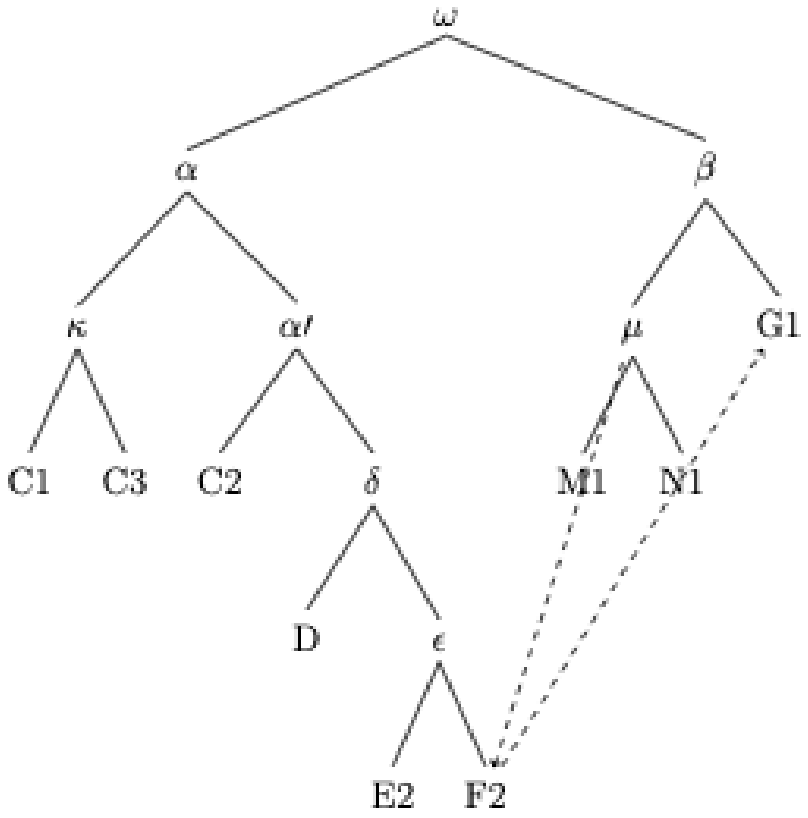


FIGURE 4 – Stemma établi à partir de la collation des variantes de la *Vie de saint Martin*

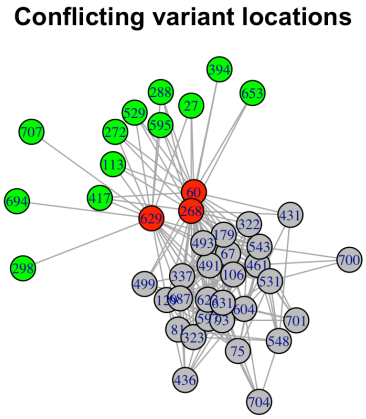


FIGURE 5 – Exemple de clusters obtenus grâce à *Stemmatology*

```
</app> <app>
  <lem>a</lem>
  <rdg wit="#F2" type="outil">en</rdg>
</app>&esp;<app>
  <lem>arche<lb/>&u-v;esqes </lem>
  <rdg wit="#G1" type="ajout">estre arcevesques</rdg>
</app>de&esp;<placeName ref="#tours">
```

FIGURE 6 – Exemple d'encodage de l'apparat critique issu de l'édition numérique de la *Vie de saint Martin*

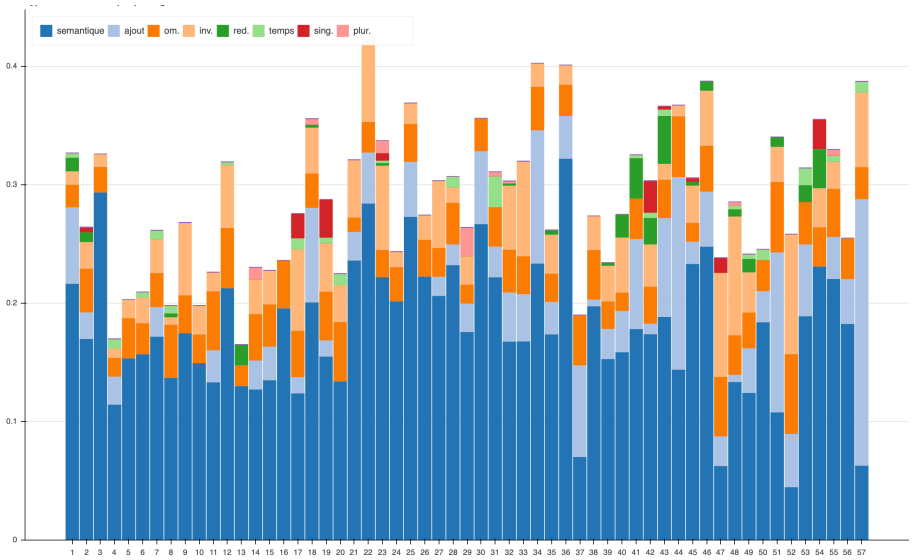


FIGURE 7 – Étude de la répartition des types de variantes à partir de la collation de la *Vie de saint Martin*

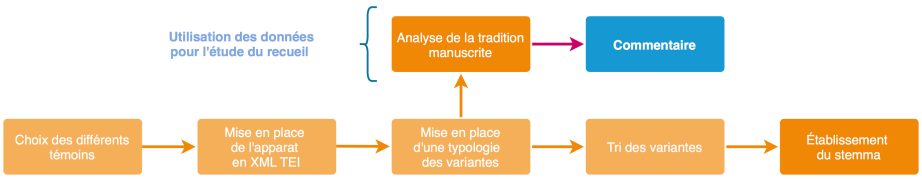


FIGURE 8 – Schéma de la réutilisation des données issues de la collation dans l'édition numérique

```
<w xml:id="t1" n="1" lemma="de" pos="PRE" msd="MORPH=empty">De</w>
<w xml:id="t2"
   n="2"
   lemma="saint"
   pos="ADJqua"
   msd="NOMB.=s|GENRE=m|CAS=r|DEGRE=p">seint</w>
<w xml:id="t3"
   n="3"
   lemma="Martin"
   pos="NOMpro"
   msd="NOMB.=s|GENRE=m|CAS=r">Martin</w>
```

FIGURE 9 – Exemple d'encodage des annotations linguistiques

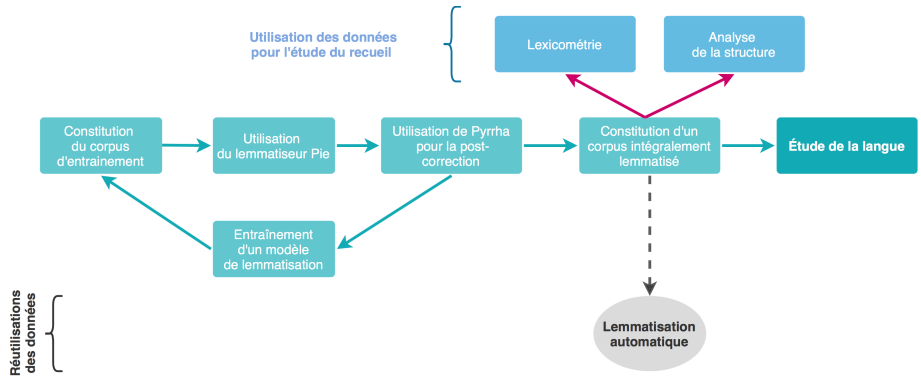


FIGURE 10 – Schéma de la réutilisation des données de lemmatisation dans l'édition numérique

— title : Exemple d'encodage des annotations linguistiques keywords : données, linguistique type : image lang : fr link : images/09LemmatisationCode.png date : 2021 source : auteur priority : lowpriority —

— title : Schéma de la réutilisation des données de lemmatisation dans l'édition numérique keywords : données, lemmatisation type : image lang : fr link : images/10Lemmatisation.png date : 2021 source : auteur priority : lowpriority —

— title : Exemple d'encodage des développements d'abréviations keywords : données, lemmatisation type : image lang : fr link : images/11ChoiceCode.png date : 2021 source : auteur priority : lowpriority —

```
<!ENTITY etilde-em '<choice><abbr>ẽ</abbr><expan>e<ex>m</ex></expan></choice>'>
<!ENTITY etilde-en '<choice><abbr>ẽ</abbr><expan>e<ex>n</ex></expan></choice>'>
<!ENTITY etilde-ez '<choice><abbr>ẽ</abbr><expan>e<ex>z</ex></expan></choice>'>
```

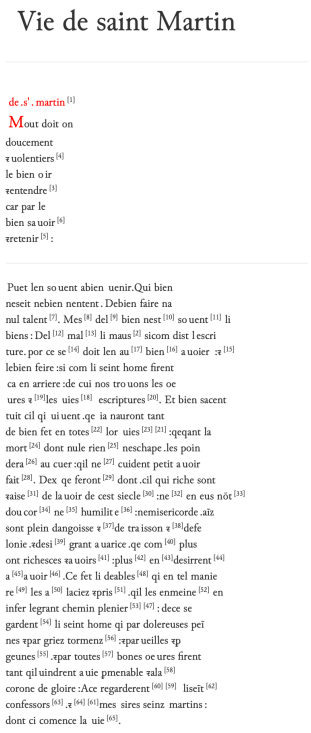
FIGURE 11 – Exemple d'encodage des développements d'abréviations

Table 2 – Sélection des scores TF-IDF les plus hauts pour les Vies de saint Benoît, saint Jérôme et saint Alexis

Vie de saint Benoît		***Vie de saint Jérôme***		***Vie de saint Alexis***	
Terme	*score*	*Lemme*	*score*	*Lemmes*	*Score*
saint	0.312	asne	0.262	saint	0.240
frere	0.106	lïon	0.225	nostre	0.187
moine1	0.073	saint	0.198	emperëor	0.125
abeie	0.069	chamoil	0.166	cité	0.103
nostre	0.067	frere	0.140	fil2	0.101
chose	0.065	nostre	0.119	vie1	0.096
oraison	0.062	pasture1	0.097	maison	0.083
eglise	0.052	marchëant	0.094	pere	0.080
avenir	0.049	creche	0.073	povre	0.079
tu	0.046	chose	0.071	tu	0.075
foie2	0.038	eglise	0.063	chartre1	0.067
maniere	0.035	busche	0.062	mïemement	0.067
comencier	0.035	ebrieu	0.059	o4	0.065
lieu	0.034	ues	0.053	feme	0.063
dïable	0.034	caldeu	0.051	voiz	0.059
tantost	0.033	latin	0.050	tresque1	0.059
parole	0.033	comander	0.050	chose	0.057
ensemble	0.033	porte1	0.049	jovencel	0.057
prestre	0.033	clochier2	0.046	ainsi	0.055

Table 3 – Sélection des scores les plus hauts pour l'indice de spécificité de Lafon dans les Vies de saint Benoît, saint Jérôme et saint Alexis

Vie de saint Benoît		***Vie de saint Jérôme***		***Vie de saint Alexis***	
Terme	*spécificité*	*Lemme*	*Spécificité*	*Lemmes*	*spécificité*
Beneoit	Infini	Jerome	81,99	Alexis	73,99
saint	181,03	asne	36,98	Eufamianus	60,92
frere	90,11	lïon	31,66	que2	36,97
orison	57,99	chamoil	20,97	saint	18
moine1	47,43	que2	17,44	nostre	17,51
abeie	39,27	frere	15,47	Rome	14,89
il	38,18	Bethleem	14,88	et	11,41
si	35,15	saint	13,82	vie1	9,81
ome	32,16	pasture1	12,33	maison	9,32
Maurus	29,54	marchëant	10,58	emperëor	9,1
xfoie2	20,39	si	9,88	povre	8,57
avenir	18,32	il	9,77	cité	8,3
un	17,82	nostre	8,34	si	8,25
Zalla	17,29	creche	7,65	seignor	7,93
chose	16,86	Rome	7,36	maisniee	7,33
eglise	15,84	come1	7,03	mïemement	7,24
nostre	15,8	latin	6,11	fil2	6,91
costume	14,33	busche	5,99	voiz	6,81
Theoprobus	12,86	ues	5,86	pere	6,5
roche3	11,8	ebrieu	5,71	chartre1	6,46



Vie de saint Martin

de .s'. martin [1]
Mout doit on
doucement
ꝛ uolentiers [4]
le bien o ir
ꝛentendre [3]
car par le
bien sa uoir [6]
ꝛretenir [5] :

Puet len so uent abien uenir.Qui bien
neseit nebien nentent . Debien faire na
nul talent [7]. Mes [8] del [9] bien nest [10] so uent [11] li
biens : Del [12] mal [13] li maus [2] sicom dist l escri
ture. por ce se [14] doit len au [17] bien [16] a uoier :ꝛ [15]
lebien feire :si com li seint home firent
ca en arriere :de cui nos tro uons les oe
ures ꝛ [19]les uies [18] escriptures [20]. Et bien sacent
tuit cil qi ui uent .qe ia nauront tant
de bien fet en totes [22] lor uies [23] [21] :qeqant la
mort [24] dont nule rien [25] neschape .les poin
dera [26] au cuer :qil ne [27] cuident petit a uoir
fait [28]. Dex qe feront [29] dont .cil qui riche sont
ꝛaise [31] de la uoir de cest siecle [30] :ne [32] en eus nôt [33]
dou cor [34] ne [35] humilit e [36] :nemisericorde .aïz
sont plein dangoisse ꝛ [37]de tra isson ꝛ [38]defe
lonie .adesi [39] grant a uarice .qe com [40] plus
ont richesces ꝛa uoirs [41] :plus [42] en [43]desirrent [44]
a [45]a uoir [46] .Ce fet li deables [48] qi en tel manie
re [49] les a [50] laciez ꝛpris [51] .qil les enmeine [52] en
infer legrant chemin plenier [53] [47] : dece se
gardent [54] li seint home qi par dolereuses peï
nes ꝛpar griez tormenz [56] :ꝛpar ueilles ꝛp
geunes [55] .ꝛpar toutes [57] bones oe ures firent
tant qil uindrent a uie pmenable ꝛala [58]
corone de gloire :Ace regarderent [60] [59] liseït [62]
confessors [63] .ꝛ [64] [61]mes sires seinz martins :
dont ci comence la uie [65].

FIGURE 12 – Transcription graphématique de la *Vie de saint Martin*, d'après le manuscrit fr. 412 de la Bnf, fol.103r

— title : Transcription graphématique de la Vie de saint Martin, d'après le manuscrit fr. 412 de la Bnf, fol.103r keywords : transcription, manuscrit type : image lang : fr link : images/12TranscriptionGraph.png date : 2021 source : auteur priority : lowpriority —

— title : Transcription normalisée de la Vie de saint Martin, d'après le manuscrit fr. 412 de la Bnf, fol.103r keywords : transcription, manuscrit type : image lang : fr link : images/13TranscriptionNorm.png date : 2021 source : auteur priority : lowpriority —

— title : Schéma de la réutilisation des données de transcription dans l'édition numérique keywords : transcription, données type : image lang : fr link : images/14Transcription.png date : 2021 source : auteur priority : lowpriority —

— title : Schéma de la réutilisation des données dans le cadre de l'édition numérique des Seint Confessor keywords : édition, données type : image lang : fr link : images/15Workflow.png date : 2021 source : auteur priority : lowpriority —

Vie de saint Martin

1. De seint Martin [1] mout doit on doucement et volentiers [4] le bien oïr et entendre [3], car par le bien sa voir [6] et retenir [5]
[fol. 103b] puet l'en so vent a bien venir. Qui bien ne seit ne bien n'entent de bien faire n'a nul talent [7]. Mes [8] del [9] bien nest [10] so v
ent [11] li biens, del [12] mal [13] li maus [2] si com dist l'Escriture. Por ce se [14] doit l'en au [17] bien [16] a voier et [15] le bien feire, si com li
seint home firent ça en arriere de cui nos tro vons les oe vres et [19] les vies [18] [es] escriptures [20]. Et bien sacent tuit cil qi vi vent qe ja n'
auront tant de bien fet en totes [22] lor vies [23] [21] qe, qant la mort [24] dont nule rien [25] n'eschape les poindera [26] au cuer, q'il ne [27]
cuident petit a voir fait [28]. Dex ! Qe feront [29] dont cil qui riche sont et aise [31] de l'a voir de cest siecle [30], ne [32] en eus n'ont [33] dou çor
[34] ne [35] humilit é [36] ne misericorde, ainz sont plein d'angoisse et [37] de tra isson et [38] de felonie et de si [39] grant a varice qe com [40] plus
ont richesces et a voirs [41], plus [42] en [43] desirrent [44] a [45] a voir [46] ? Ce fet li deables [48] qi en tel maniere [49] les a [50] laciez et pris [51] q'il
les enmeine [52] en infer le grant chemin plenier [53] [47]. De ce se gardent [54] li seint home qi par dolereuses peinnes et par griez tormenz [56]
et par veilles et par geunes [55] et par toutes [57] bones oe vres firent tant q'il vindrent a vie parmenable et a la [58] corone de gloire. A ce
regarderent [60] [59] li seint [62] confessors [63] et [64] [61] messires seinz Martins dont ci comence la vie [65].

FIGURE 13 – Transcription normalisée de la *Vie de saint Martin*, d'après le manuscrit fr. 412 de la Bnf, fol.103r

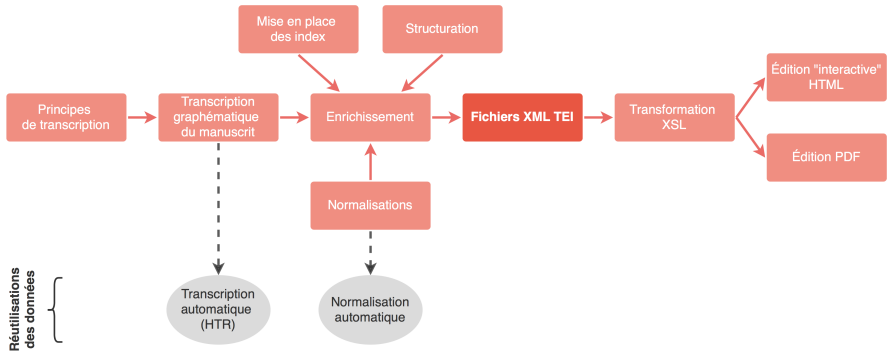


FIGURE 14 – Schéma de la réutilisation des données de transcription dans l'édition numérique

Table 4 – Table des abréviations et de leurs développements dans le recueil des Seint Confessor

Abr.	dév.	Nb	Termes
7	et	4764	et
ꝑ	par	275	aparut, departement, departi, departie, departies, departir, departir, departirent, departirent, departiroit, departisist, departisist, departoit, espartirent, par, pardon, parfete, parla, parler, parlions, parloit, parlé, parmenable, parmi, parole, paroles, pars, part, partie, parties, partir, partout, parvenir, parvenue, parz, parler, partie, partirent, partout
9	con	247	aconplir, confessor, congié, conmanda, conmandas, conmandast, conmande, conmandement, conmandemenz, conmander, conmanderent, conmanderoit, conmandez, conmandoit, conmandé, conmant, conmanz, conme, conmence, conmencement, conmencier, conmencierent, conmencié, conment, conmença, conmençames, conmençast, conmunement, conneües, conneüssent, connissoit, connoistre, conpagnie, conpaignons, conselle, consomee, conterai, contre, contree, controverei, conté, conurent, convient, encontre, raconte
ẽ	en	194	aeuroient, alerent, amerent, amoient, aoroient, argent, arrestoient, assemblerent, assiduelment, assistrent, avoient, baptizierent, bien, biens, boivent, ceens, cenz, chastement, citoien, comencierent, comença, condenpnabitur, conmencierent, conment, conurent, covient, croient, deguerpeïssent, demanderent, dementieres, demorerent, descendoit, descendre, desirroient, diemence, distrent, doivent, dolenz, donerent, empense, emporteroient, en, enorter, ensanglente, entendi, entendoient, entendoit, ententive, entiegent, entrefirent, envoierent, erent, esgardoient, esmurent, estoient, fesoient, firent, foïssent, furent, fussent, genz, governement, griement, habitent, hantent, indulgense, jovenceaus, jovencel, lendemein, longement, mandement, meintient, mendiz, mengier, mengoit, menroient, mistrent, moveroient, noient, oceïssent, ocistrent, Orient, paien, paiens, partirent, pensoient, peüssent, porpensa, portoient, prendoit, prendre, prendroient, prensignast, prensigniez, pristrent, proierent, puent, reclaiment, rendez, rendi, repristrent, respondirent, retiegnent, retrestrent, revellierent, rieng, riens,, siens, sovent, tenpore, tesmoignent, tesmoigneroient, trenchier, trentiesme, trestrent, troverent, veillierent, venoient, vent, ventreil, venz, veoient, vindrent, virent, voelent, voient, volent, volentiers
ĩ	in	120	ainz, ausint, aussint, avint, Benjamin, ceint, certeinnement, compleindre, destreins, einsint, einz, estreinz, larrecin, loing, Martin, Martins, mein, meins, meintenant, meintenir, meintenissent, meinz, peinnes, pleins, revint, seint, seinte, seintismes, seinz, tesmoing, venins, ving, vint, voisins
õ	on	84	adonc, adont, anoncierent, avons, barons, beneïçon, bons, compaignon, compaignons, condenpnabitur, confusïon, conmença, conpaignie, conpaignie, conseil, conter, contes, conté, devotion, dont, dragon, environnee, Jethron, meson, mesons, mon, monde, mons, monsigneur, montaigne, Noiron, non, noncierent, nonmee, nonmé, ont, orissons, oroisons, parfondece, pissonniaus, proions, puissons, raconte, raconter, raconterei, racontes, resoignons, respondi, respondoit, respons, son, sont, temptation, translation, trencherons, vision
ꝑ	per	72	apercevoir, apercevoit, aperceü, aperçut, apertement, desesperance, empereor, empereors, emperere, empereres, esperit, esperiz, pere, peres
ã	an	59	anssamblé, avant, creance, demanda, demandai, demorance, devant, doutance, enfant, ensanble, estrange, France, grant, marcheanz, meintenant, portanz, preecant, ramembrance, repentanz, sanblance, sanz, semblant, serjant, tant, tramblant

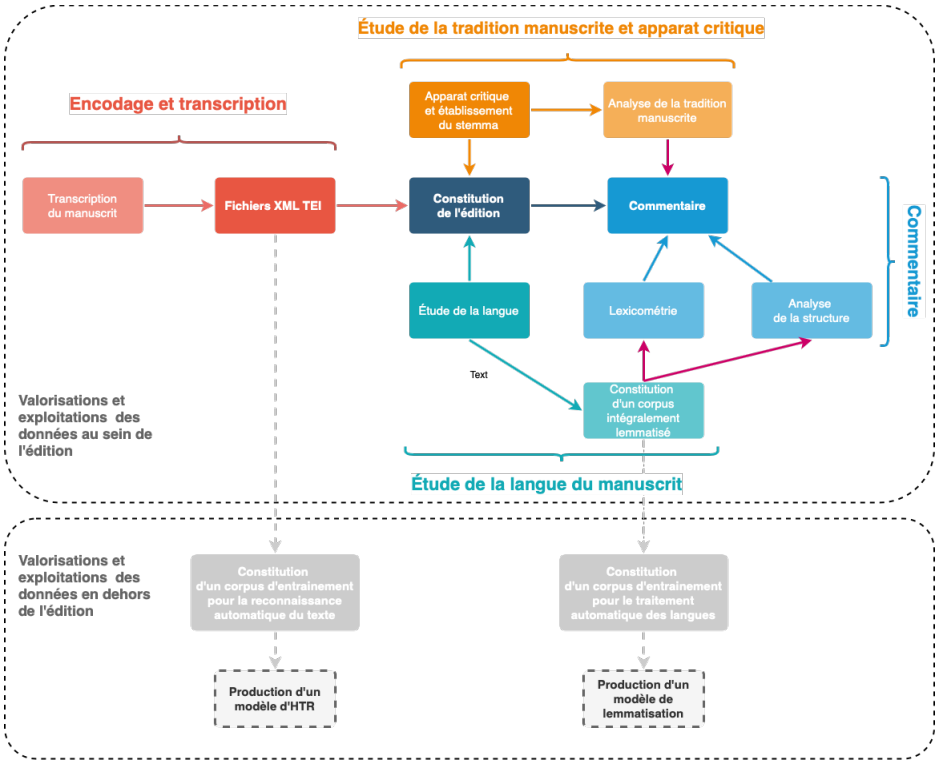


FIGURE 15 – Schéma de la réutilisation des données dans le cadre de l'édition numérique des *Seint Confessor*

5 Pour approfondir la lecture

Jean-Baptiste Camps, Thibault Clérice, Ariane Pinche, *Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis*, 2019, halshs-03044086

6 Biographie

Ariane Pinche (ORCID : 0000-0002-7843-5050) est docteure en langue et littérature médiévales et postdoctorante à l'École nationale des chartes. Elle s'intéresse tout particulièrement à l'édition de numérique et a remporté le prix Fortier de la meilleure communication jeune chercheur lors de la conférence *Digital Humanities* 2019 à Utrecht avec ses deux collègues J. B. Camps et T. Clérice pour la communication Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis. Aujourdhui, ses intérêts de recherche se portent tout particulièrement sur la confection de corpus médiévaux pour l'entrainement dHTR (Handwritten Text Recognition).
Voir le CV en ligne