

UNIVERSITÉ PARIS, SCIENCES & LETTRES

Malamatenia Vlachou-Efstathiou

Diplômée de Licence en Lettres Classiques

Diplômée de Master en Latin

**Éditer les manuscrits grammaticaux glosés :
solutions numériques face aux défis
traditionnels**

Le cas du *Voss.Lat.O.41* d'Eutychès

Mémoire de première année du master
« Humanités Numériques »

juin 2022

Résumé

En règle générale, les manuscrits grammaticaux glosés constituent un défi pour les chercheurs qui souhaitent les éditer. Leur caractère hétérogène reflété dans les spécificités structurelles de leur *mise-en-page*, ainsi que dans les multiples couches d'annotations et de variations orthographiques, font de leur modélisation une tâche difficile. Les outils numériques, comme en témoignent des projets récents, offrent une variété de solutions qui se révèlent particulièrement efficaces et donc nécessaires pour leur édition, du HTR et les vocabulaires contrôlés jusqu'au XML-TEI et la visualisation des données. En prenant le *Vossianus Latinus O41* du *de uerbo* d'Eutyches comme cas d'étude, nous tâcherons de mettre en œuvre un pipeline semi-automatique vers une édition documentaire multifonctionnelle.

Mots-clés : paléographie latine ; Grammatici Latini ; manuscrits glosés ; Eutychès ; humanités numériques ; HTR ; apprentissage machine ; ALTO ; XSL ; SegmOnto ; XML-TEI

Informations bibliographiques : Malamatenia Vlachou-Efstathiou, *Éditer les manuscrits grammaticaux glosés : solutions numériques face aux défis traditionnels : Le cas du Voss.Lat. O.41 d'Eutychès*, mémoire de master 1 « Humanités Numériques », dir. [Cécile Conduché, Peter A. Stokes], Université Paris, Sciences & Lettres, 2022.

Abstract

Generally, glossed grammatical manuscripts pose a significant challenge to researchers that wish to edit them. Their heterogenous character reflected in the structural specificities of their *mise-en-page*, alongside often multiple layers of annotations and orthographical variation, make their modeling a difficult task. Digital tools, as recent projects attest, offer a variety of solutions that reveal particularly efficient and thus necessary for their edition, from HTR and controled vocabularies to XML-TEI and the visualisation of the data. With Eutyches' *Vossianus Latinus O41* of the *de uerbo* as a case study, we'll try to apply a semi-automatised pipeline towards a multifunctional documentary edition.

Keywords : latin paleography ; Grammatici Latini ; glossed manuscripts ; Eutyches digital humanities ; HTR : machine learning ; ALTO ; XSL ; SegmOnto ; XML-TEI

Bibliographic Information : Malamatenia Vlachou-Efstathiou *Editing glossed grammatical manuscripts : digital solutions to traditional challenges : The case of the Voss.Lat. O.41 of Eutyches*, mémoire de master 1 « Humanités Numériques », dir. [Cécile Conduché, Peter A. Stokes], Université Paris, Sciences & Lettres, 2022.

Remercîments

Avant la présentation du projet, je tiens à remercier brièvement tous ceux qui ont contribué à l'élaboration de ce mémoire, à quelque étape et de quelque façon que ce soit. Leur aide s'est révélée extrêmement précieuse tout au long de cette année, alors que je me lançais progressivement dans le domaine des Humanités Numériques.

Tout d'abord, je voudrais remercier mes deux directeurs de recherche. D'une part Mme Cécile Conduché, qui m'a fait découvrir l'œuvre d'Eutychès et ses défis, a partagé avec moi son travail et a fait confiance à mes capacités face à un manuscrit inédit. De l'autre Mr. Peter Stokes, qui a accepté de m'encadrer et qui a été patient dès le début alors que je me familiarisais lentement avec les spécificités de mon jeu de données et toutes les possibilités que le numérique avait à offrir à ma recherche. Sur le même plan, je remercie chaleureusement Mme Cécile Lanéry de l'IRHT qui m'a appris tout ce que je connais sur la paléographie latine et dont la passion pour cette discipline je partage et j'admire.

En même temps, je suis très reconnaissante aux professeurs de l'ENC qui ont bien voulu m'accompagner dans mon travail et qui ont pris le temps, dans les moments de détresse, d'écouter patiemment mes problèmes et mes idées et de m'aider à trouver des solutions. Je me réfère plus particulièrement à Mr. Thibault Clérice, Mme Lucence Ing et Mr. Chahan Vidal-Gorène sans l'orientation desquels ce projet n'aurait pas été le même. Un sincère merci va également à l'équipe de *Gallicorpora*, Mme Ariane Pinche, Mr. Simon Gabay, et mes collègues, avec qui j'ai eu la chance de travailler et d'apprendre, notamment de m'avoir fait découvrir le potentiel d'eScriptorium et de Segmonto.

Je voudrais remercier aussi mes camarades et amis, d'avoir partagé avec moi cette année académique intense, de m'avoir fait me sentir la bienvenue malgré mon origine étrangère, et de m'avoir fait part de leurs propres projets, passions et connaissances. Sans eux, tout ce processus n'aurait pas été aussi agréable et constructif. Enfin, un énorme merci à mon père et à ma sœur qui m'ont toujours soutenue quoi qu'il arrive, et ont cru en moi quand je ne le faisais pas.

Table des matières

Résumé	ii
Abstract	ii
Table des matières	v
1 Introduction	I
1.1 Eutychès et le <i>De uerbo</i>	I
1.2 La typologie des manuscrits grammaticaux latins	3
1.2.1 Les tableaux d'exemples	3
1.3 <i>Quaestiones</i> d'édition	5
1.3.1 Description du jeu de données - MsDesc	8
2 Un pipeline semi-automatique	II
2.1 Aquisition des données : HTR	13
2.2 Transformation des données	19
2.2.1 XSL	19
2.2.2 XML-TEI	20
2.2.3 Choix de l'encodage	22
2.2.4 Questions de normalisation/lemmatisation	29
2.3 Visualisation des données	30
2.3.1 Mouvements d'annotation	31
2.3.2 Typologies générales vs. spécifiques	34
3 Limites et perspectives	39
3.1 Un modèle de SegmOnt(ation)	39
3.2 Balisage automatique du couple lemme/glose	41
3.3 Nouvelles perspectives - Portée du projet	43
3.4 Conclusions générales	44
Table des figures	51

Chapitre I

Introduction

1.1 Eutychès et le *De uerbo*

On place le grammairien Eutychès (*alias* Eutex, Eutychius) à Constantinople au milieu du VI^e siècle¹. Disciple de Priscien de Césarée (dont il définit en tant que *meus preceptor*²) , sur le plan de la doctrine grammaticale, l'oeuvre d'Eutychès constitue le premier témoin de sa posterité. Auteur de deux traités grammaticaux, le premier, à savoir le *De aspiratione*, qui nous est parvenu seulement via la tradition indirecte, à savoir dans le chapitre IX « De aspiratione », du *De orthographia* de Cassiodore, traite le *b* graphique en la langue latine. C'est seulement le deuxième traité, à savoir le *De uerbo* qui jouit d'une tradition directe, et qui constitue le sujet de notre recherche.

Dans cet essai grammatical articulé en deux livres³, Eutychès répond à la demande de *Craterus*, un de ces disciples, qui se demandait comment assigner les verbes latins au bon modèle de conjugaison à partir de leur forme de base (la première personne du singulier du présent de l'indicatif à l'époque). L'auteur se propose d'y indiquer les critères formels permettant de reconnaître à quelle conjugaison appartiennent les verbes latins. Après l'exposé de quelques principes généraux - par exemple que les verbes incohéatifs relèvent de la troisième conjugaison- Eutychès passe en revue une liste finie des suffixes de dérivation verbale et nominale. Et cela du fait qu'un verbe peut être soit dérivé et alors selon son suffixe appartient à un modèle de conjugaison, soit c'est une forme primaire, alors c'est sa famille lexicale qui sert comme indice. Il procède avec une analyse méticuleuse des *finalitates verborum*, c'est-à-dire des sons ou groupes de sons qui précèdent la désinence -o/-ou et indique de quel type flexionnel chaque *finalitas* est caractéristique. Cette analyse peut être caractérisée comme discursive parfois réduite aux schématismes, d'un style simple et aride, d'ailleurs abondamment illustrée de listes d'exemples. Dans cette analyse, des nombreux passages d'auteurs

1. Ce sous-chapitre constitue une compilation des informations dans Valeria Lomanto, « Eutiche », dans *Encyclopédia virgiliana*, 1985, Cécile Conduché, « La mise en page d'Eutychès », dir. François Roudaut (, 2019), p. 51-68 et James EG Zetzel, *Critics, compilers, and commentators : An introduction to Roman philology, 200 BCE-800 CE*, 2018p.298

2. Henricus Keil, *Grammatici latini ex recensione Henrici Keilii.-Lipsiae, BG Teubner 1857-1880*, t. 1-5, 1857, vol5, p.456

3. annoncé par l'auteur lui même dans : *GL*, vol.5, p.447, 12-14 *ad discernendas pertinens coniugationes, duobus libellis inclusi. Quorum prior obseruationibus instruitur generalibus, alter inditio finalitatis, spetiales exequitur regulas.*

classiques sont purement illustratifs : en effet, Eutyches ne soulève pas de doutes ni ne discute de cas controversés, mais décrit avec soin les phénomènes de la langue latine qui n'impliquent pas des disparités entre l'usage commun et l'usage littéraire. Pour ces raisons il a été critiquée par les philologues qui n'y reconnaissait pas une valeur intrinsèque. Cette approche reflète la tendance générale aux travaux grammaticaux jusqu'au milieu du 20^e siècle, où, en raison de l'intérêt croissant pour la linguistique et l'histoire de la pensée linguistique, l'étude des grammairiens latins a connu un renouveau dans un cadre plus scientifique⁴.

Ce ne fut qu'en 1974 que Jeudy Colette⁵ a recensé la tradition directe du *De uerbo* en identifiant vingt-neuf témoins manuscrits⁶ du *De uerbo*, que ça soit complets ou fragmentaires, s'étendant du VIII^e au XI^e siècle.

Quant à la postérité de l'oeuvre, témoin de son succès en Haut Moyen-Âge, la diffusion d'Eutychès en Occident a commencé assez rapidement, peut-être déjà de son vivant, en tous cas avant la fin du VI^e siècle, où l'on date l'oeuvre de Cassiodore. Son traité a été lu et utilisé jusqu'au XI^e siècle, ce dont témoigne à la fois la tradition directe et indirecte du traité, ainsi que les commentaires dont il a été l'objet. Quant à l'étendue géographique de son influence, *De uerbo* doit indubitablement sa renommée aux Irlandais, les deux témoins les plus anciens⁷ étant en minuscule irlandaise. Depuis le début du IX^e siècle, on constate la présence de ce manuel également en Angleterre et au Continent, dans tout l'empire carolingien, centrée essentiellement dans trois régions : en France (surtout dans le Nord et l'Est), en Bavière et dans la région du lac de Constance, dans des milieux monastiques comme Luxueil, Corbie, Saint-Gall et Fleury entre autres.

À part sa diffusion manuscrite, la fortune du *De verbo* est également documentée par son utilisation indirecte dans les grammaires du haut Moyen Âge. C'est Bengt Löfstedt, éditeur d'une bonne partie de ces dernières⁸, qui a attiré l'attention sur la présence d'Eutychès parmi leurs sources. Cette tradition indirecte remonte au moins un siècle avant les plus anciens manuscrits conservés, à la seconde moitié du VII^e siècle, avec le traité anonyme intitulé *Ad Cuimnanum*, du nom de son dédicataire. Par la suite, durant toute la période carolingienne, les grammairiens ont continué à exploiter le *De verbo* dans la rédaction de leurs propres traités, à savoir Sedulius Scottus et Rémi d'Auxerre, dont on parlera plus tard à l'occasion de leur mise-en-page. Une tradition aussi riche a relancé l'attention des chercheurs contemporains, sans pour autant avoir encore donné lieu à une

4. Viven Law, Louis Holtz, Mario DeNonno et Paolo de Paolis ne sont que quelques-uns des noms qui ont mené des recherches pionnières dans ce domaine.

5. Colette Jeudy, *Les manuscrits de l'”Ars de uerbo” d'Eutychès et le commentaire de Rémi d'Auxerre*, 1974

6. Le site MMDC mentionne sur le codex composite Leiden, UB : ms. BPL 154 : 1, ff. 001-037 « This text is a compilation of excerpts from Priscian's *Institutiones grammaticae* and from Eutyches' *Ars de verbo* », information d'ailleurs nulle part mentionnée dans la littérature. Le ms. en question n'étant pas numérisé, cette information reste à vérifier.

7. le premier copié au VIII^e à Bobbio Naples, lib naz. Lat.2 (mahlheureusement passé au réactif), le deuxième en Irlande même, avec des gloses irlandaises : Il s'agit des deux fragments BnF.Paris 10400 et 11411.

8. Bengt Löfstedt, *Der hiberno-lateinische Grammatiker Malschanus*, Uppsala, 1965 puis id., « Zu Tatwines Grammatik », *Arctos* 7, 1972, p. 47-65.

nouvelle édition.

1.2 La typologie des manuscrits grammaticaux latins

La tradition manuscrite des textes grammaticaux se distingue par deux caractéristiques inhérentes, unanimement prêtes : une mise en page complexe et la présence de gloses. Louis Holtz, érudit et théoricien majeur des manuscrits grammaticaux, dans l' introduction à son article « La typologie des manuscrits grammaticaux latins »⁹ résume l'importance des caractères codicologiques globaux qui en font l'essence :

[...]type d'écriture, mise en page, recours aux abréviations, décoration. Tous ces éléments, dans la mesure où ils participent d'un système donné, ont leur importance pour rendre compte aussi bien de la survie des textes transmis que pour expliquer la forme particulière dans laquelle il nous sont parvenus. [...] ces éléments proprement codicologiques, qu'un chercheur, dès lors qu'il a recours à ce document qu'est le manuscrit, ne peut éluder, même s'il n'a en vue qu'une édition critique, car ils sont susceptibles d'éclairer ou d'amplifier les données auxquelles lui donne accès le collationnement du texte lui-même.

1.2.1 Les tableaux d'exemples

Par exemple, plusieurs témoins présentent des exemples de conjugaisons verbales en colonnes ordonnées ou non. Si ces dernières facilitent la lecture et par conséquent, la mémorisation des exemples et ne servent pas simplement d'inventaire, elles pourraient également être représentatives et évocatrices de l'usage concret du document copié, tout en donnant une idée des témoins disponibles aux copistes et sur lesquels circulaient les œuvres grammaticales pendant le Moyen Âge. Vivien Law dans son article « From aural to visual »¹⁰, a appelé à analyser cet usage à un niveau culturel plus profond, dans un bref. elle voyait dans cette « présentation tabulaire la marque du caractère avant tout oral de la grammaire et liait la représentation visuelle, l'usage des diagrammes dans la grammaire progrès de l'analyse morphologique¹¹.

Et si Remigio Sabbadini, responsable d'une collation partielle du codex glosé Milan, Ambr. B71sup.¹², avait raison de déplorer l'impossibilité de restaurer le texte d'Eutychès dans sa forme originale et la futilité d'amasser des collations, du fait que chaque copiste modifiait ou changeait

9. Louis Holtz, « La typologie des manuscrits grammaticaux latins », *Revue d'histoire des textes*, 7-1977 (1978), p. 247-269, p.247

10. Vivien Law, « From Aural to Visual : Medieval representations of the word », *Grammar and Grammarians in the early Middle Ages* (, 1997), p. 250-259

11. cf. également le chapitre « Memory and Structure of Grammars » de Vivien Law dans le premier volume de M De Nonno, P de Paolis et L Holtz, « Manuscripts and tradition of grammatical texts from antiquity to the Renaissance » (, 2000) pour une analyse plus détaillée.

12. Remigio Sabbadini, *Opere minori : Classici e umanisti da codici latini inesplorati*, t. 87, 1995, p.76-83

cette liste à sa guise en transcrivant horizontalement les colonnes verticales et vice versa, cette observation n'est pas en vain pour la fortune du texte tout court. Pour le cas d'Eutyche, un examen approfondi¹³ des tableaux ordonnés des exemples dans la tradition manuscrite, comparés aux artes influencés par *De uerbo*, démontre l'importance d'une telle analyse. Par conséquent, cet aspect ne doit donc assurément pas être négligée par notre étude non plus.

1.2.1.1 Le couple indissociable lemma-glose

Vu le nombre de gloses présentes dans la tradition manuscrite du *De uerbo*, il convient d'en donner une brève définition et d'élucider leur nature et leur fonction. Comme c'est le cas pour de nombreuses œuvres classiques influentes, les espaces interlinéaires et marginaux d'un manuscrit des *artes* sont généralement couverts de gloses. L'étude de ce paratexte riche et complexe éclaire la réception des *Artes* et la tradition même des études grammaticales dans l'Antiquité tardive et au Moyen Âge, et a fait l'objet d'une attention accrue dans la recherche récente¹⁴. Etant donnée la variété des gloses apposées sur le texte grammatical (simple synonymes, notes qui élucident le texte, observations grammaticales) celui-ci offre un exemple manifeste de la relation dynamique entre le texte et le paratexte.

Ayant évoqué ces deux termes, une distinction entre le texte et le paratexte, auquel les gloses appartiennent, s'impose. En règle générale, la glose apporte une clarification du sens de certaines phrases et même de certains mots (*non solum sententiam sed etiam verba attendit*¹⁵). Le texte, en revanche, est le livre de l'auteur sans autres remarques explicatives (*textus est liber sine littere vel sententie ex positione*)¹⁶ dont les mots sont autant de lemmes potentiels. Cependant, la frontière même entre texte et paratexte parfois s'avère poreuse. En raison de cette confusion, de multiples discussions ont été menées et de multiples définitions données sur le sens des marginalia, des notes, des commentaires et des gloses¹⁷. En fait, on peut trouver d'innombrables « niveaux de paratextualité » selon la localisation et le but que sert une glose. On trouve également un troisième niveau de paratexte, étroitement lié aux gloses et aux marginalia, c'est les *glossaria* à part entière. Le rôle des glossaires, était de fournir une aide à la compréhension de base de la manière dont le texte était construit. Éviter la confusion et expliquer les incertitudes devaient être les principes directeurs du glossateur à ce stade. Malgré ces difficultés occasionnelles, les gloses interlinéaires sont généralement courtes, claires et sans ambiguïté. Leur qualité d'éclairage et leur valeur éducative

13. L'étude unique qui porte sur cette question est la suivante, dotée d'un tableau comparatif des occurrences des listes ordonnées dans la tradition manuscrite : C. Conduché, « La mise en page d'Eutychès »...

14. Paolo Monella, « A digital critical edition model for Priscian », *M. Pade (ed. by), Philology Then and Now : History, Role, and New Directions [im Erscheinen]* (, 2019)

15. Pour l'évolution sémantique du terme de l'Antiquité au Moyen-Âge voir Louis Holtz, « Glossaires et grammaire dans l'Antiquité », dans *Les manuscrits des lexiques et glossaires de l'Antiquité tardive à la fin du moyen âge : Actes du Colloque international (Erice, 23-30 septembre 1994)*, 1996, p. 1-21

16. Greti Dinkova-Bruun, « Text and gloss », dans *The Oxford Handbook of Latin Palaeography*, 2020, p. 925

17. cf. particulièrement le chapitre de Adolfo Tura « Essai sur les *marginalia* en tant que pratique et documents » dans : Danielle Jacquart et Charles SF Burnett, *Scientia in margine : études sur les marginalia dans les manuscrits scientifiques du moyen âge à la renaissance*, t. 88, 2005, qui offre une vue panoramique sur l'espace paratextuel des manuscrits

sont indéniables. Louis Holtz¹⁸ remarque sur la notion des éléments textuels secondaires qu'il entend : « tout ce qui vient se greffer après coup sur le texte d'un auteur [...] qui n'ont pas d'autre raison d'être que de faciliter, de guider, d'orienter la lecture [...] bref, tout ce qui dans nos livres manuscrits n'émane pas de l'auteur lui-même [...] ». Franck Cinato se penche sur le concept et le status des gloses dans le contexte grammatical¹⁹, et propose sa propre définition *ontologique* : « toute augmentation péritextuelle qui précise ou diversifie l'information contenue dans un texte principal. » qui, de l'échelle codicologique, « documents the idea's intellectual afterlife²⁰ »

Des différents niveaux de sophistication de la glose peuvent être et sont souvent mis en jeu simultanément, même si fréquemment le processus de glose semble être façonné par les besoins intellectuels d'un lectorat particulier. En particulier lorsque les textes étaient utilisés pour l'enseignement, la complexité de leurs notations an reflétait le stade éducatif auquel ces textes étaient intégrés au programme. Dans certains manuels médiévaux, nous pouvons même trouver de multiples couches de gloses introduites indépendamment les unes des autres par différents utilisateurs du livre sur une longue période de temps. Dans de tels cas, les gloses sont perçues comme étant en dialogue non seulement avec le texte proprement dit, mais aussi avec les autres couches d'annotation. Même lorsqu'elles résultent d'une lecture et d'une étude privées, les gloses individuelles varient en difficulté et en objectif, reflétant les inclinations érudites de la personne qui les a écrites. En conséquence, la composition primaire et le commentaire secondaire sont mutuellement façonnés en une nouvelle réalité textuelle à la fois plus riche en signification et plus complexe en forme. Il n'est pas suffisant de se contenter à l'idée que les gloses et les commentaires ont été écrits uniquement pour fournir l'élucidation de textes obscurs et difficiles. L'entreprise de glossologie médiévale est bien plus que cela. Elle représente un mode de pensée, un mode d'expression et une méthode complexe d'engagement intellectuel avec la tradition littéraire et savante héritée²¹. Suivre un tel processus est important pour l'histoire des idées, la transmission et la posterité du texte et l'enseignement de la grammaire. Ceci est encore plus important lorsqu'il s'agit de la tradition multidimensionnelle d'une œuvre, depuis son texte principal et ses gloses, jusqu'aux commentaires externes et aux glossaires.

1.3 *Quaestiones* d'édition

1.3.0.1 Interet philologique d'une édition des gloses du *De uerbo*

En prenant en considération la manière dont les grammaires de l'Antiquité tardive ont été éditées, Louis Holtz n'avait pas non plus tort quand il prononça que les grandes autorités dans le champ des études grammaticales du (xix^e siècle), Keil, Wessner, Barwick, Tolkiehn, éprouvaient une certaine répulsion à l'égard des textes grammaticaux du Moyen Age. Aussi importants que soient les éléments présentés *infra*, ils étaient en même temps particulièrement déroutants pour les éditeurs

18. Louis Holtz, « Les manuscrits latins à gloses et à commentaires : de l'antiquité à l'époque carolingienne », dans *Il Libro e il testo*, dir. R. Raffaelli C. Questa, 1984, p. 139-167, p. 142

19. Franck Cinato, *Priscien glosé*, t. 41, 2015, p.187-198

20. G. Dinkova-Bruun, « Text and gloss »..., p.924

21. *Ibid.*, p.938

modernes, qui, confrontés à la réalité de l'édition imprimée, choisissaient d'ignorer les éléments structurels qui échappaient au domaine du texte principal et décontextualisaient ainsi (soit le texte des gloses et des tableaux de conjugaisons) soit les gloses du contexte (Le cas par excellence étant le *Corpus Glossariorum Latinorum* de Goetz). L'édition monumentale en 7 tomes des *Grammatici Latini* de Heinrich Keil à la fin du XIX^e siècle, reste, dans la plupart des cas, l'édition de référence pour l'étude des GL. Malgré son importance pour l'étude des GL, dans son édition toute preuve de mise en page et d'annotation des manuscrits d'origine est absente. La tendance à isoler les textes de leur contexte, fortement imposée par la composition complexe des manuscrits, prive ces témoins de leurs véritables richesse²² et potentiel (co-occurrences, variantes significatives).

Franck Cinato, dont la recherche constitue pour nous un *exemplum* méthodologique important, annonçant son propre travail sur la tradition des manuscrits glosés de Priscien de Césarée²³, résume de manière concise mais pointue l'intérêt et les enjeux d'une édition de ce type de matériel. Résumons le champs d'intérêt concernés par une édition des gloses :

1. l'histoire des théories linguistiques au Moyen Âge ;
2. l'histoire de l'enseignement : corrélations entre les préoccupations des maîtres et le contexte historique dans lequel elles s'inséraient ;
3. l'étude de la réception du texte et des transferts de connaissances : relations entre *scriptoria* à l'époque carolingienne et post carolingienne ou des courants doctrinaux dans le contexte des universités qui se trouveront étayés par un fondement solide constitué de l'ensemble des détails extraits des gloses²⁴ ;
4. une cartographie globale de l'histoire de la constitution des gloses : diffusion et répartition géographique du fonds commun et mise en évidence de la chronologie des innovations ;
5. mise en corrélation du fonds des gloses grammaticales avec les gloses bibliques : évaluation de l'influence du travail des *grammatici* dans la constitution de la culture intellectuelle médiévale

Dans le cas d'Eutychès, dont l'œuvre a jouit des commentaires à part entière pendant le haut Moyen Âge, l'impact d'une édition des gloses est encore plus immédiat sur notre perception de la postérité de l'œuvre. Il est rare que les textes grammaticaux antiques nous parviennent coupés des commentaires qui constituent leur « rénouvellement » médiévale. Le *De uerbo* doit surtout ce renouveau aux commentaires redigés par deux personnages et pédagogues identifiables, commençant

22. Elena Pierazzo et Peter A Stokes, « Putting the text back into context : a codicological approach to manuscript transcription », dans *Kodikologie und Paläographie im digitalen Zeitalter 2—Codicology and Palaeography in the Digital Age 2*, 2011, p. 397-430

23. F. Cinato et EPHE PARIS, « Perspectives offertes par un corpus électronique de gloses sur Priscien », *Eruditio antiqua*, 3 (2011), p. 131-51, p. 145-6

24. Vivien Law, chercheuse pionnière dans le domaine des GL souligne cet apport en se référant aux manuscrits de Priscien : « Many scholars besides those known to us by name jotted their learning and insights on the margins of manuscripts of the *Ars maior* and *Institutiones grammaticae*. It is high time that modern researchers began to investigate what they had to say. » V. Law, *Grammar and grammarians in the Early Middle Ages*, 1997, p. 146

par le commentaire *In Eutychem* par Sedulius Scottus²⁵ au début du IX^e siècle alors qu'il était encore en Irlande. Quelque temps plus tard, dans un tout autre style, Rémi d'Auxerre (v. 841-908) en fit aussi un commentaire, conservé sous forme des gloses marginales anonymes, aux fol. 81-97 du ms. 1470 de la Bibliothèque municipale de Rouen, reproduites aussi dans le codex BnF, lat. 7499²⁶.

Quant à la typologie des manuscrits qui conservent plus précisément le commentaire de Rémi d'Auxerre, celle-ci comprend à la fois des gloses interlinéaires et des commentaires continus en *catena* organisés par lemmes et présente une mise en page étayée et hiérarchisée. Lui-même ayant fait des « collations » entre plusieurs codices²⁷, ce codex hybride comporte le texte, les gloses et le commentaire de Rémi en marge, fait partie intégrale de la tradition à la fois du texte principal et des gloses y associées, étant donnée qu'il avait accès aux matériels aujourd'hui perdus, qui complètent la tradition manuscrite de l'œuvre²⁸. Dans ces conditions, il serait imprudent de séparer en deux lots les opuscules d'un manuscrit grammatical et de n'accorder qu'une médiocre attention aux commentaires médiévaux et au lien étroit qu'ils entretiennent avec l'œuvre originale. Ce serait se priver de nombreuses sources d'information sur le texte antique lui-même, sa transmission, la qualité de sa recension, le lieu où il a été copié, informations qui précisément font souvent défaut pour d'autres types de textes. La manière dont les lecteurs contemporains et tardifs de ces textes ont réellement donné un sens à leur doctrine et l'ont appliquée en dehors du domaine immédiat de la grammaire sont des domaines d'étude qui commencent tout juste à attirer l'attention. Comment les grammairiens du cinquième et sixième siècles étaient lus par les grammairiens des dixième et onzième - ce que les historiens de la linguistique aimeraient savoir.

Pour conclure cette partie théorique et pour en resumer les points centraux, à l'encontre des œuvres littéraires, les grammaires latines ont été copiées au cours des siècles non pour leur « valeur intrinsèque » mais en tant que manuels scolaires, selon l'utilité concrète et l'intérêt qu'elles présentaient en termes d'apprentissage du latin. Pour cette raison, selon le niveau d'érudition et les besoins pratiques, les copistes/compilateurs, eux-mêmes souvent érudits, proposent des synonymes, des gloses, des corrections et des commentaires en marge et souvent reproduisent des tableaux d'exemples et d'exceptions. Ceci rend la typologie des manuscrits grammaticaux multidimensionnelle, étayée et, en quelque sorte, personnalisée dans chaque témoin, tout en gardant un rapport

25. édité par Scottus, S., & Löfstedt, B. (1980). In *Donati Artem Minorera*; In *Priscianum*; In *Eutychem*. *Tijdschrift Voor Filosofie*, 42(1).

26. C. Jeudy, *Les manuscrits de l'"Ars de uerbo" d'Eutychès et le commentaire de Rémi d'Auxerre...*, p. 434-436. On précise que le commentaire de Rémi d'Auxerre reste largement inédit J. E. Zetzel, *Critics, compilers, and commentators: An introduction to Roman philology, 200 BCE-800 CE...*, p. 352. La seule édition à ce jour est partielle : Manitius, « *Rernigusscholien* », *Münchener Museum für Philologie des Mittelalters und der Renaissance* II, p. 101-108.

27. C. Jeudy, *Les manuscrits de l'"Ars de uerbo" d'Eutychès et le commentaire de Rémi d'Auxerre...*, p. 435. En bon philologue, il s'intéresse aux diverses leçons des manuscrits qu'il a sous les yeux et aux divergences d'interprétation : *Caluesco* (448, 28) : « *Alii codices habent caluisco id est caluere incipio, id est decipere, et alii caluesco, id est caluus fieri incipio* » (Paris 7499, fol. 73 v).

28. Louis Holtz, « *La typologie des manuscrits grammaticaux latins* ...», p. 258 section « *De la glose anarchique au commentaire organique* » et p. 260 « *Fluidité des systèmes* ».

d'interdependance entre eux. Nous avons dit des manuscrits grammaticaux que ce sont des instruments de travail et leur mise en page incarne leur rôle pédagogique. En même temps, en tant que *custodes Latinitatis*, les grammairiens, soit via la doctrine linguistique qu'ils présentent, soit via le renvoi – par le biais de citations – aux œuvres littéraires et grammaticales perdues, sont très précieux dans l'aperçu de l'enseignement de la grammaire, la littérarité et l'évolution linguistique pendant l'Antiquité tardive. Leur popularité pendant le Moyen Âge, qui se traduit en de nombreuses copies de leurs œuvres, témoigne de l'état de l'apprentissage du latin sur le continent européen à cette période-là.

Même si la littérature secondaire est assez conséquente et s'améliore en termes de rigueur scientifique, il n'existe toujours pas d'éditions numériques natives agrégeant la totalité des témoins et proposant une transcription « dynamique » sur les GL.. La préparation de l'édition critique de Priscien (projet PAGES) à l'Université Sapienza²⁹ de Rome et l'édition critique numérique des gloses du premier livre des *Etymologiae* d'Isidore de Séville d'Evina Steinová³⁰ font donc figure d'exception qui ouvrent la voie pour des chercheurs/éditerus à venir. En effet, plusieurs chercheurs regrettent l'absence d'une édition critique du *De uerbo* d'Eutychès qui permettrait une étude plus objective et fondée de son œuvre vis à vis du réseau du corpus des GL. Notre projet s'inscrit à cette initiative des éditions natives numériques pour les Grammairiens Latins, en concentrant son intérêt à Eutychès. Mettre un oeuvre une stratégie d'édition qui prendra en considération les spécificités de ces documents, permettra l'exploitation de la richesse intrinsèque ,récemment valorisée, qu'ils possèdent. Le fait que :

1. Plusieurs manuscrits glosés restent inédits
 2. La plus part des œuvres édites par Keil ont besoin d'une nouvelle édition critique³¹
- ne fait cette tâche plus impérative que jamais.

1.3.1 Description du jeu de données - MsDesc

Parmi la liste des manuscrits récencés par Jeudy Colette, nous avons choisi en tant que jeu de données, après discussion avec Mme. Conduché, le manuscrit inédit Leiden, Vossianus Latinus *in octavo* 41³², que Keil a écarté en tant que *codex descriptius* lors de son édition. Il s'agit d'un *codex* composite, qui accueille presque exclusivement des ouvrages grammaticaux, le *De uerbo* d'Eutychès et les *Etymologiae* d'Isidore de Séville. D'abord une brève description : VLO41 est un manuscrit en parchemin du dernier quart du (ix^e siècle) qui compte 65 folios paginés en haut à droite. La feuille

29. P. Monella, « A digital critical edition model for Priscian »...

30. Evina Steinová et Peter Boot, « The glosses to the first book of the Etymologiae of Isidore of Seville : a digital scholarly edition » (, 2021)

31. La richesse de travail que ce corpus réserve se traduit par le fait que malgré la parution constante des nouvelles éditions, en 2018 seulement 40/103 textes inclus dans le corpus de Keil ont été remplacés par une édition postérieure de KEIL J. E. Zetzel, *Critics, compilers, and commentators : An introduction to Roman philology, 200 BCE-800 CE...*, p.160, n.5

32. Dorénavant VLO41.

de garde (1r) consiste en une liste de la correspondance de Grégoire le Grand.

Quant à sa composition codicologique, les folios sont disposés en 8 cahiers de taille non homogène, à savoir : IV (2-9) + V (10-19) + IV (20-27) + III (28-33) (fin de l'oeuvre d'Eutychès) + 4 IV (34-65). Le codex manque un feuillet déchiré entre les folios 9v et 10r ce qui coincide au passage du quaternion (à l'origine un quinon) au deuxième quinon. Par contre le sens de lecture n'est pas perturbée et la main reste la même jusqu'au folio 22r. On note ici que les 33 premiers folios du manuscrit qui comportent l'oeuvre d'Eutychès correspondent à la deuxième unité codicologique³³. Des marques explicites de possession d'*Isaac Vossius* et de Paul Petau apparaissent au folio 2r, et le manuscrit, d'après une comparaison avec le codex *VL O37*, doit avoir appartenu au XII^e siècle à l'abbaye de Fleury (DeMeyer³⁴ : *33v scriptura evanuit; insuper custos quidam bibliothecae Floriacensis (?) xii litteris maioribus ab imo ad summum codicis indicem scripsit: Liber Ysidori iunioris cu(m) euticio grammatico; cf. ad cod. O.37*).

Des initiales ornées figurent dans les folios 2r, 3r, 8v (celle-ci à l'encre verte et jaune) et un beau « O » majuscule entouré des croix dans le folio 19v, marquant le début du deuxième livre. Des initiales capitales de taille normale mais accentuées (dépassant ou non d'une ligne de hauteur) marquant potentiellement le début d'un paragraphe se trouvent partout dans le folio, et sont parfois des ajouts postérieurs³⁵. Des motifs fonctionnels en forme de croix marquent le début/la fin des chapitres (lorsque les initiales ne le font pas) dans les folios 2r, 3r, 3v, 4v, 5v. Dans certains folios, l'encre jaune sert de surlieur pour des mots (3r, 4r) ou même pour des colonnes entières (4v, 5r). Deux dessins ont été conservés légèrement effacés dans les folios 5v (une tête) et 10v (une figure tournant le dos au texte), et un troisième semble avoir été effacé dans le folio 29r. Des marqueurs des chapitres visiblement postérieurs en chiffres romains surlignés à l'encre verte, allant de VIII (folio 13v) à XII (folio 17v), y figurent également.

Un total de quatre mains principales et huit mains de glosateurs participent à la confection du manuscrit, qui s'étendent sur plus d'un siècle (DeMeyer considère, d'après l'écriture, que les mains de gloseurs appartiennent entièrement au Xe siècle). Le manuscrit comporte plusieurs essais de plume, dont une série de neumes de la séquence "Virgo in uiolata", le mot in uiolata figurant en-dessous des neumes. Une souscription d'un certain copiste Concordus au folio 25r et une incertaine/effacée (*mibi nom̄ ++ ind̄ erit/ uer?um*) dans la marge du folio 9v sont également présentes.

Le texte principal est écrit strictement en 25 lignes, en respectant la réglure, sauf pour le folio 4r qui comporte 27 lignes après l'ajout postérieur d'une partie du texte. Sur la mise en page des folios 4r, 4v et 5r : Le folio 4r présente à la fin 3 colonnes de taille égale en guise de tableau

33. Karel A De Meyer, *Codices vossiani latini*, t. 16, 1973, p. 80

34. *Ibid.*, p. 81.

35. Précisément dans les folios 5r, 5v, 6v, 10r, 12v, 13v, 14r, 15r, 16r, 17r, 19r, 23v, 26r, 29v, 30r.

d'exemples qui occupe la deuxième moitié du folio. Le folio 5r présente 4 petites colonnes à la fin en guise de tableau d'exemples qui occupent le 1/3 de la page. Une partie du texte est incluse dans une petite cellule en haut à gauche. Le folio 5r présente la plus grande hétérogénéité avec 4 colonnes de taille inégale parsemées dans la page en perturbant le sens de lecture. Selon sa typologie, VLO41 constitue un cas particulièrement intéressant qui correspond au type Γ . Plus particulièrement il est question d'une grammaire glosée, d'une collection propre à un témoin manuscrit unique où les gloses se trouvent à proximité immédiate du texte ; elles sont le résultat des travaux de glossateurs qui se sont succédés durant une période plus ou moins longue.

Sur la base des considérations théoriques susmentionnées, et disposant d'un cas d'étude pratique présentant plusieurs des caractéristiques qui s'avèrent difficiles à gérer, nous avons développé un pipeline englobant, qui repose sur des outils numériques. De l'acquisition des données à leur structuration et visualisation, plusieurs outils ont été mis en œuvre afin de permettre leur mise en valeur. Les chapitres qui suivent visent à élucider le processus que nous avons suivi, et à mettre en avant les avantages que les outils numériques offrent pour l'étude du jeu de données spécifique, tout en évoquant les limitations de notre recherche et les pistes d'amélioration.

Chapitre 2

Un pipeline semi-automatique

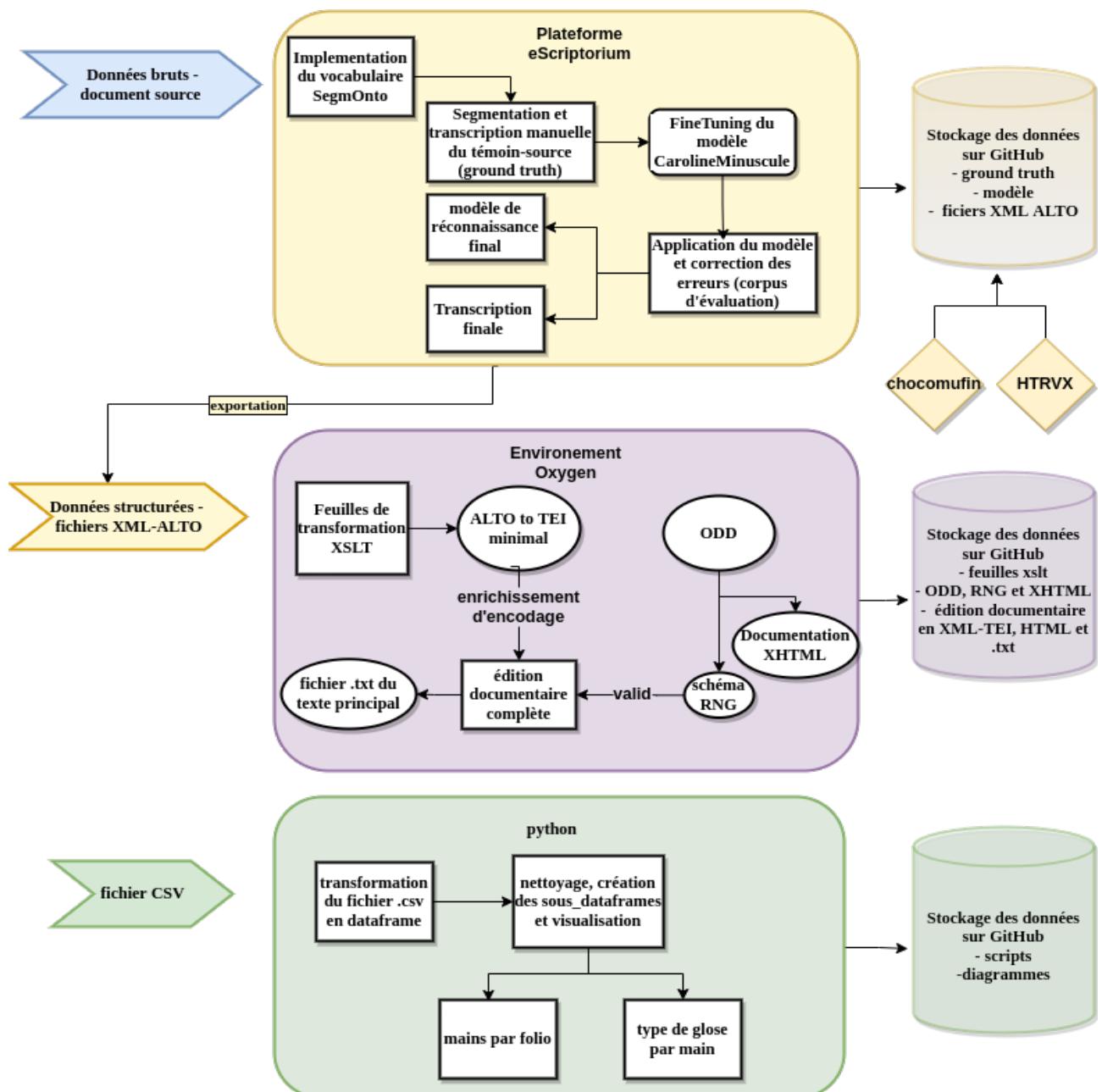


FIGURE 2.1 – Modélisation du pipeline suivi. Trois niveaux de traitement successifs, de la transcription à la structuration et la visualisation des données.

Le pipeline suivant a été pensé en prenant en compte tous les aspects dont : les besoins, les objectifs, les outils, les enjeux techniques, les résultats attendus, le type d'analyse sur textes extraits, et les résultats voulus. Idéalement, l'objectif de ce projet est le développement d'un pipeline englobant qui prend en compte les spécificités de tous nos témoins.

2.1 Aquisition des données : HTR

Il existe de nombreux logiciels d'OCR (Reconnaissance Optique de Caractères) et d'HTR (Reconnaissance automatique de structure et d'écriture manuscrite) : Il s'agit de la conversion d'un texte imprimé, écrit ou inscrit, en un texte encodé par une machine par un logiciel de reconnaissance automatique de texte qui analyse une image numérisée pour en extraire le texte. Certains sont propriétaires et d'autres sont gratuits, et ils ont tous des spécificités qui les rendent plus adaptés à un projet plutôt qu'à un autre. En ce qui concerne le logiciel utilisé pour le projet sur Eutychès, nous avons utilisé *kraken*¹, via son interface *eScriptorium*², un outil d'analyse de mise en page et d'HTR fondé sur de l'apprentissage profond (IA). Ce choix s'explique par deux raisons principales. Tout d'abord, il s'agit d'un logiciel libre avec une documentation étendue (que ça soit des tutoriels en ligne ou la documentation dans leur dépôt [GitLab](#)). Pour ce qui est de plus, le personnel de l'ENC et d'Inria a activement et extensivement travaillé sur plusieurs projets en utilisant *kraken* et *eScriptorium*, ce qui a facilité considérablement à résoudre des problèmes que nous avons rencontrés au fur et à mesure de notre recherche, qu'ils soient d'ordre technique ou méthodologique.

2.1.0.1 Analyse de la mise en page complexe / Complex Layout Analysis

La première étape vers l'acquisition de nos données, dans notre cas les différentes couches de texte contenues dans la copie numérique des folios de VLO41, est leur segmentation. La segmentation constitue le processus qui consiste à décomposer l'image entière en sous-parties pour les traiter ensuite. Dans le cas du VLO41, la segmentation de l'image est effectuée dans l'ordre suivant : segmentation au niveau de la zone et segmentation au niveau de la ligne. *Kraken* propose déjà un modèle par défaut pour la segmentation des pages écrites de gauche à droite, entraîné sur un grand ensemble de données, qui traite assez bien les lignes principales de la page, à condition que la qualité de l'image soit bonne. Néanmoins, compte tenu des spécificités susmentionnées et de l'hétérogénéité considérable de la mise en page, la segmentation de la page avec le modèle intégré n'était pas dans tous les cas capable de gérer la localisation inattendue de l'information, comme par exemple plusieurs lignes interlinéaires ou marginales écrites en petits caractères. Il n'était pas rare non plus pour le modèle de confondre le texte avec le paratexte, en mélanger les lignes principales avec les lignes interlinéaires ou marginales. Face à cette gestion incomplète et afin de conserver la mise en

1. <https://kraken.re/master/index.html>

2. Benjamin Kiessling, Robin Tissot, Peter Stokes et Daniel Stökl Ben Ezra, « eScriptorium : An open source platform for historical document analysis », dans *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, IEEE, 2019, t. 2, p. 19-19

page et la hiérarchie inhérente aux différents niveaux, une approche manuelle et personnalisée a été adoptée pour la segmentation de chaque page. Cette étape de *Complex Layout Analysis* consiste à détecter les zones d'intérêts : c'est-à dire les zones de textes, la pagination, les colonnes, les reclames etc. ; éventuellement, il s'agit de différencier les niveaux sémantiques distincts qui coexistent dans une image.

Dans ce cadre, la stratégie suivante a été mise en place : chaque ligne principale a été notée comme telle du début à la fin, en constituant une unité sémantique. Par contre, il n'en a pas été de même pour l'espace interlinéaire. En effet, une ligne interlinéaire était séparée en autant de lignes que de gloses présentées en son sein. Et cela parce que la disposition spatiale au sein de l'espace interlinéaire ne correspond pas à une ligne où existe un flux continu de sens. Les gloses interlinéaires, en tant qu'extensions du texte principale, n'étaient pas conçues pour être lues l'une après l'autre de manière linéaire. Cette approche a été mise en œuvre a priori comme un moyen de donner aux gloses un statut d'autonomie tenant compte également de l'extraction et exploitation des données. Alors dans notre système, une ligne dans l'espace interlinéaire vaut pour une glose, celle si s'afférent à un lemme. En revanche, comme c'est le cas avec les manuscrits d'usage, la situation n'est pas aussi simple qu'une modélisation simpliste en fait. Au moins pour les interlignes, dans de rares occasions (4 au total) les gloses s'étalent sur deux lignes au lieu d'une, rendant inopérante la règle $1\text{ligne}=1\text{glose}$. Étant donné cette petite exception (4/ 1050) à la règle générale, veillant à l'uniformité de nos données, nous avons fait le choix éditorial de transférer la moitié inférieure vers la moitié supérieure de la glose, en ajoutant le signe "/" pour signaler le saut de ligne existant dans le manuscrit. Par exemple, pour la glose *multitudo ho/minum* sur *uulgu*s qui se trouve dans la ligne 21 du folio 11r, la barre verticale oblique indique le saut à la ligne. Cela constitue une des limites de notre segmentation, notamment en ce qui concerne la validité de la vérité de terrain³, par contre indispensable pour la manipulation des données.

SEGMONTO

Dans la même veine, une opération complémentaire a été également adoptée, en faisant usage des vocabulaires contrôlés. Un vocabulaire contrôlé est un lexique dont le but est de permettre l'organisation des connaissances pour optimiser la recherche d'informations. Le vocabulaire contrôlé est utilisé dans les schémas d'indexation par sujet, les thésaurus et les taxonomies et nécessite l'utilisation de termes prédéfinis qui ont été présélectionnés par le concepteur du vocabulaire. Afin de repérer les différentes zones du document et le type de lignes présentes dans la page ainsi que de les caractériser d'un point de vue codicologique, nous avons décidé d'implémenter le vocabulaire contrôlé SegmOnto⁴. SegmOnto est né du besoin pour une application informatique d'un vocabu-

3. La vérité de terrain consiste en « des ensembles de données annotées et corrigées de manière à fournir au modèle des paires composées d'une part d'une image ou d'une portion d'image (entrée) et d'autre part de l'annotation attendue (sortie), qui peut être des coordonnées dans le cas de la segmentation ou un ensemble de caractères pour la transcription. Les performances des modèles dépendent certes de l'architecture neuronale mise en place, mais aussi de la qualité et de la quantité de vérité de terrain fournies lors de l'apprentissage. » Définition tirée de Alix Chagué, Thibault Clérice et Laurent Romary, « HTR-United : Mutualisons la vérité de terrain ! » (, 2021)

4. Simon Gabay, Jean-Baptiste Camps, Ariane Pinche et Claire Jahan, « SegmOnto : common vocabulary and

laire commun limité, une ontologie basé sur les normes existantes, pour la description et l'analyse de la mise en page de documents, allant de la catégorisation du contenu à la reconnaissance de texte. SegmOnto aborde principalement le cas des manuscrits, ce qui favorise davantage son utilisation. Cette pratique répond aux besoins de plus en plus pressants dans le domaine de l' HTR. En effet, avec l'apparition d'analyseurs de mise en page efficents et d'interfaces faciles à utiliser, le besoin de modèles efficaces de segmentation augmente, tout comme le besoin de grandes quantités de données, basée sur l'agrégation de documents hétérogènes. Pour cela, les chercheurs doivent à se mettre d'accord sur un vocabulaire commun limité, et partager des pratiques communes pour faciliter l'interopérabilité de leur vérité de terrain. En s'adhérant à ce cadre scientifique et en reconnaissant les avantages d'une méthode de description -dotée d'une empreinte numérique sur l'export en format ALTO- pour les documents hétérogènes à plusieurs niveaux, nous avons implementé SegmOnto pour VLO4I.

En pratique, la caractérisation des zones et des lignes est très adaptée au cas du VLO4I, étant donnée les doubles contraintes de la mise en page où l'ordre de lecture est perturbé, et les trois niveaux d'information que l'ajout des gloses impose. Prenons l'exemple d'application le plus manifeste : Considérant qu'eScriptorium fait une lecture horizontale les lignes en allant de gauche à droite, les colonnes, introduisant une lecture verticale, sont ignorées⁵. Par conséquent, tracer manuellement les différentes « zones de lecture » avec SegmOnto est une condition indispensable. La page se divise ainsi dans les zones suivantes selon l'information qu'elle portent :

- *NumberingZone* pour la foliation ;
- *MainZone*, ou eventuellement *MainZone#1,#2* et ainsi de suite pour la zone qui comporte le texte principale ainsi que les gloses ;
- *MarginTextZone* pour les gloses marginales ;
- *MusicZone* pour les neumes ;
- *GraphicZone* pour les dessins.

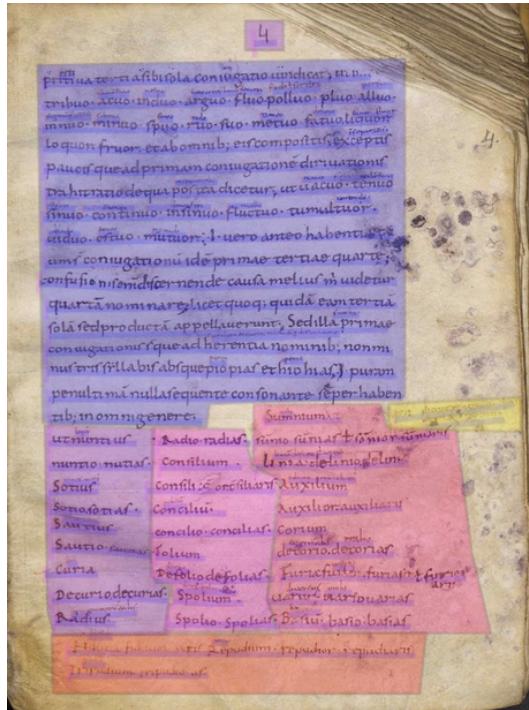
Il en va de même pour les deux types de lignes. Grâce à SegmOnto, les deux niveaux d'information, les lignes principales (marquées en tant que *DefaultLine*) et les interlignes (marquées en tant qu' *InterlinearLines*) peuvent être distinguées et caractérisées, ne laissant aucune ambiguïté sur la nature de la ligne en question, tout en facilitant l'extraction et l'exploitation des données. En outre, afin d'assurer un contrôle qualité de notre export lors du stockage des information sur GitHub, le logiciel *HTRVX*⁶, développé par Thibault Clérice et Ariane Pinche a été mise en place vérifie le schéma XML et l'utilisation de l'ontologie Segmonto pour la segmentation.

practices for analysing the layout of manuscripts (and more) », dans *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*, 2021

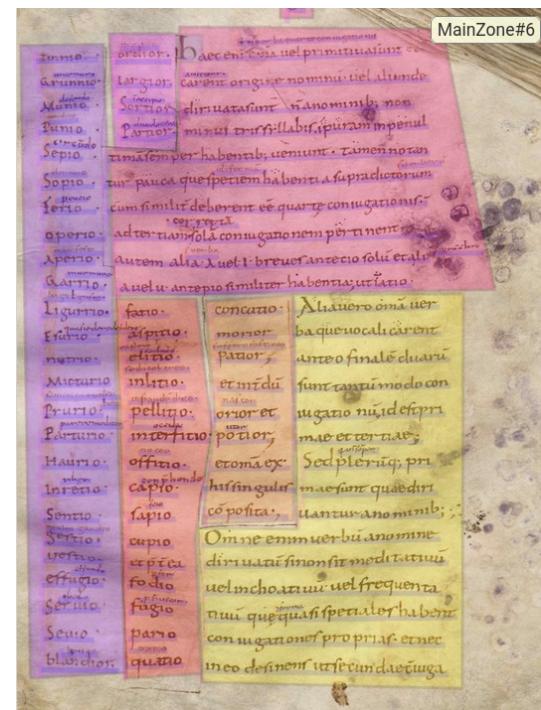
5. Et pour citer Mr. Vidal-Gorène dans un récent entretien pour BULAC (cf. supra n.9) : « Ainsi pourrais-je obtenir un CER* de 0%, mais si les lignes ne sont pas lues dans le bon ordre, est-ce pour autant utile ? »

6. T. Clérice et A. Pinche, *HTRVX, HTR Validation with XSD*, version 0.0.1, sept. 2021, doi : [10.5281/zenodo.5359963](https://doi.org/10.5281/zenodo.5359963)

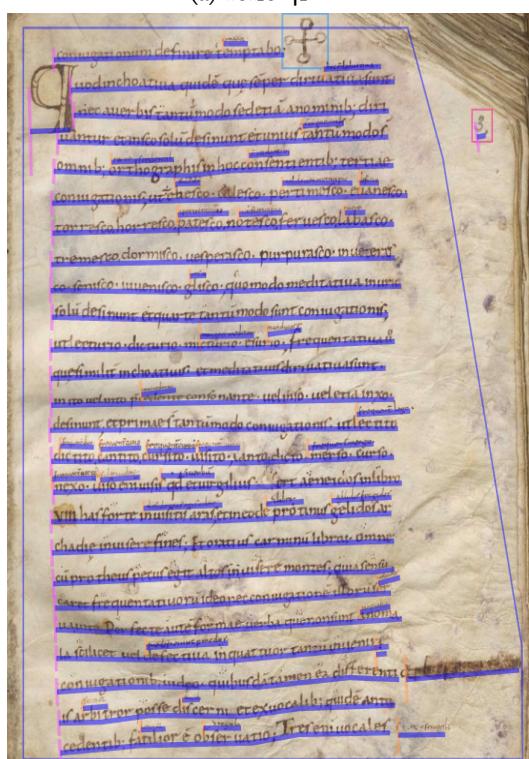
Des exemples de gestion du zonage et des lignes sont présentés dans la Figure 2.2 ci-dessous :



(a) folio 4r



(b) folio 5r



(c) folio 3r

FIGURE 2.2 – (a) et (b) Disposition des zones SegmOnto afin de rétablir le sens de lecture. (c) DefaultLines en rose, InterlinearLines en orange.

2.1.0.2 Reconnaissance

Jusqu'à très récemment, les anciennes écritures manuscrites telles que la minuscule caroline ont été considérées comme imperméables aux outils de transcription automatique comme kraken. Certes, il était toujours possible d'entraîner des modèles sur un ensemble de données, mais un modèle générique « clé en main » capable d'atteindre une précision élevée générale pour plus qu'un manuscrit n'était pas facile à mettre en place⁷. Depuis 2019 et l'avancement des moteurs OCR qui gèrent les écritures manuscrites, cette difficulté a été partiellement résolue, en permettant une plus grande interoperabilité entre modèles HTR.

Un tel modèle « clé en main » nous avait été fourni par Mr. Thibault Clérice⁸ qui, à l'époque, atteignait le taux de précision de près de 92%. Tout de même, une fois appliqué, celui-ci nécessitait une grande campagne de post-correction pour notre manuscrit. En effet, en raison de l'estompage considérable de l'encre, le modèle n'était pas efficace pour l'ensemble des pages. Et bien sûr, cela ne concerne que le texte principal et en aucun cas les gloses de petite taille qui l'entourent. Il est important de noter que l'efficacité de la reconnaissance des caractères et des mots à proprement parler dépend directement du travail réalisé sur la segmentation de la page. En d'autres termes, même si le modèle de reconnaissance des caractères est performant, une segmentation incorrecte ou incomplète empêchera tout résultat satisfaisant. Pour toutes ces raisons, la plus grande partie de la transcription a été faite manuellement, dans l'optique de *fine-tuner* ou plutôt de personnaliser le modèle fourni pour qu'il réponde à nos besoins⁹. 80% du jeu de données, a été réservé pour le *fine-tuning* et 20% pour l'évaluation. La présence de, au total, huit mains distinctes dans le même jeu de données offre une variété précieuse qui empêche une surapprentissage totale sur une des mains et l'impossibilité d'une reutilisation directe du modèle de reconnaissance. Par contre, dans la sélection de la vérité de terrain les pages fortement annotées, où les masques des lignes interlinéaires se superposent avec ceux des lignes principales à tel point que la vérité de terrain est *contaminée*, ont été écartées. Le modèle final, d'un taux de 98,4% d'accuracy¹⁰, a été employé pour la transcription des trois pages du glossaire BnF.lat.14087 (corpus de test) qui a nécessité une post-correction minimale.

Afin de vérifier la qualité de notre lecture, au moins pour ce qui est du texte principal, nous

7. Sur la complexité que la minuscule caroline pose et les facteurs qui entrent en jeu lors du mélange des modèles différents voir l'article : Brandon Hawk, Antonia Karaisl et Nick White, « Modelling Medieval Hands : Practical OCR for Caroline Minuscule » (, 2018)

8. Pour la vérité de terrain de ce modèle voir. <https://github.com/rescribe/carineminuscule-groundtruth>. Le protocole de transcription s'aligne à 100% avec nos propres normes de transcription, qui visent une approche autant que possible graphématisque.

9. « Fine tuning ou spécialisation progressive : Il s'agit, dans le domaine de l'intelligence artificielle, d'adapter un modèle préexistant à un jeu de données spécifique à la tâche visée. » L'ensemble du vocabulaire spécialisé HTR et une description minutieuse de la chaîne de traitement se trouve dans l'excellent rapport rédigé par Noémie Lucas, dans le cadre d'une résidence de recherche à la BULAC en 2020-2021 soutenue et publié par le GIS, un libre accès en ligne <http://www.bulac.fr/node/2491>

10. disponible ici : <https://github.com/malamatenia/Eutyches/tree/main/data/models>

avons comparé notre transcription avec celle de Keil¹¹. Cela nous a permis de valider certaines de nos intuitions en cas de doute, de corriger certaines de nos erreurs et de mettre en évidence les endroit où le témoin s'écarte de l'édition de référence. Par contre, la lecture des presque 1050 gloses, n'a pas été toujours facile. La petite taille, le caractère informel de leur redaction et l'usure/moisissure du manuscrit n'ont pas facilité notre tâche. Dans le cas de doute, nous avons mobilisé plusieurs stratégies. En premier lieu, les autres témoins glosés¹² nous sont servis comme base comparative contre laquelle nous avons pu vérifier la lecture, notamment pour ce qui est des synonymes. Dans le cas où une étude coisée n'est pas possible, absence de correspondances entre témoins, des fouilles exhaustives dans la Database of Latin Dictionaries (DLD) et dans la Library of Latin Texts de Brepolis ont élucidé les propos des glosateurs¹³. Cette méthode nous a permis une lecture aussi fiable que possible des gloses, avec bien sûr peu de *loci* vraiment *desperati*.¹⁴

En ce qui concerne les normes de transcription, le texte a été transcrit aussi fidèlement que possible, avec une intervention minimale de notre part¹⁵. Les caractères diacritiques, les signes spéciaux¹⁶ et les abréviations et la ponctuation médiévales ont été conservés éléments précieux pour les collations éventuelles pour l'établissement du texte, nécessitant toujours une nouvelle édition¹⁷. Par contre, une intervention a été nécessaire pour la gestion des espaces entre mots dans les premières pages qui présentent des caractéristiques d'une *semi-continua*.

Finalement, comme il était le cas pour la segmentation, lors du stockage des données, le logiciel *Chocomufin*¹⁸ a été mis en place. Il s'agit d'un outil permettant de normaliser l'utilisation des caractères spéciaux issus de *mufi*¹⁹ et de vérifier la cohérence des fichiers XML. Il cible principalement le traitement des manières trop diverses de transcrire les données médiévales (allographiques et graphématisques) tout en conservant des informations telles que les abréviations. La validation *chocomufin* confirme la compatibilité de notre vérité de terrain avec le projet CREMMA²⁰ permettant une interoperabilité des données et leur réutilisation pour l'entraînement des modèles de reconnaissance automatique.

11. H. Keil, *Grammatici latini ex recensione Henrici Keilii.-Lipsiae, BG Teubner 1857-1880...*

12. Notamment le BnF.lat.7498, le glossaire BnF lat. 14087 et le BL MS. Auct. F. 4. 32 avec qui VLO41 partage une bonne partie des gloses et dont l'écriture, très lisible ne laisse pas de doutes sur la lecture.

13. Je tiens à remercier aussi Mme. Conduché pour la lecture de la glose *reicio* sur *repudium*, en bas du folio 4r., une lecture pas évidente à cause de l'usure du manuscrit

14. Tous les passages restitués sont méticuleusement signalés dans le fichier XML-TEI à l'aide de l'élément <supplied> qui permet d'enregistrer aussi le taux de certitude d'une lecture proposée.

15. Une approche purement graphématisque a servi, entre autres, à une vérité de terrain propre pour le *fine-tuning* du modèle de base.

16. Sauf un signe, qui ressemble à un *s* tironien et qui sert d'abréviation du *scilicet* qui introduit des gloses explicatives, a été transcrit en tant que *s*

17. Aux chantiers par Mme. Conduché.

18. qui vaut pour CHaracter Ocr COordination for MUFI iN textsT. Clérice et A. Pinche, *Choco-Mufin, a tool for controlling characters used in OCR and HTR projects*, version 0.0.4, sept. 2021, doi : [10.5281/zenodo.5356154](https://doi.org/10.5281/zenodo.5356154)

19. The Medieval Unicode Font Initiative : <https://mufi.info/m.php?p=mufi>

20. <https://github.com/HTR-United/CREMMA-Medieval-LAT>

ALTO

Une fois le texte segmenté, annoté et transcrit, il reste de décider le format qui se prête le mieux à l'exploitation des données structurées. Entre format .txt, PAGES et ALTO qui sont proposés par eScriptorium, le format ALTO est le plus adapté aux besoins du projet. Il va sans dire que le format .txt brut n'est pas du tout adéquat, du fait qu'il ne respecte pas ni reflète l'hierarchie interne de la mise en page et du contenu, sans parler du fait qu'il est illisible en raison de la confusion des lignes. En revanche, le format XML-ALTO (Analysed Layout and Text Object)²¹, qui est un standard XML permettant de rendre compte de la mise en page physique et de la structure logique d'un texte transcrit par reconnaissance optique de caractères (OCR) est, donc, essentiel. De plus, il est adapté à la conservation à long terme des données issues de la conversion OCR et permet une réutilisation et transformation des données en des formats différents, ce qui constitue précisément notre prochaine étape.

2.2 Transformation des données

2.2.1 XSL

Le format ultime que nous souhaitons atteindre c'est une édition documentaire en XML-TEI. Pour y arriver, il est impératif de transformer la sortie de notre transcription, à savoir les fichiers XML-ALTO en XML-TEI²². A ce but, et afin de mettre en valeur l'annotation SegmOnto, nous avons choisi d'utiliser des feuilles XSL²³. A partir de la structure des fichiers ALTO et en appliquant des règles propres à une transformation vers un format XML-TEI valide, nous avons pu obtenir le squelette²⁴ d'un fichier TEI, avec toutes les informations codicologiques concernant la foliation, les lignes principales, les interlignes, et les commentaires en marge annotées avec SegmOnto. La Figure 2.3 représente une modélisation de cette transformation ALTO vers TEI.

21. <https://www.loc.gov/standards/alto/>

22. Juliette Janes, A. Pinche, C. Jahan et S. Gabay, « Towards automatic TEI encoding via layout analysis », dans *Fantastic future* 21, 2021

23. Accessibles dans : <https://github.com/malamatenia/Eutyches/tree/main/XSL>

24. https://github.com/malamatenia/Eutyches/blob/main/XSL/out/VL041_ALTO2TEI_minimal.xml

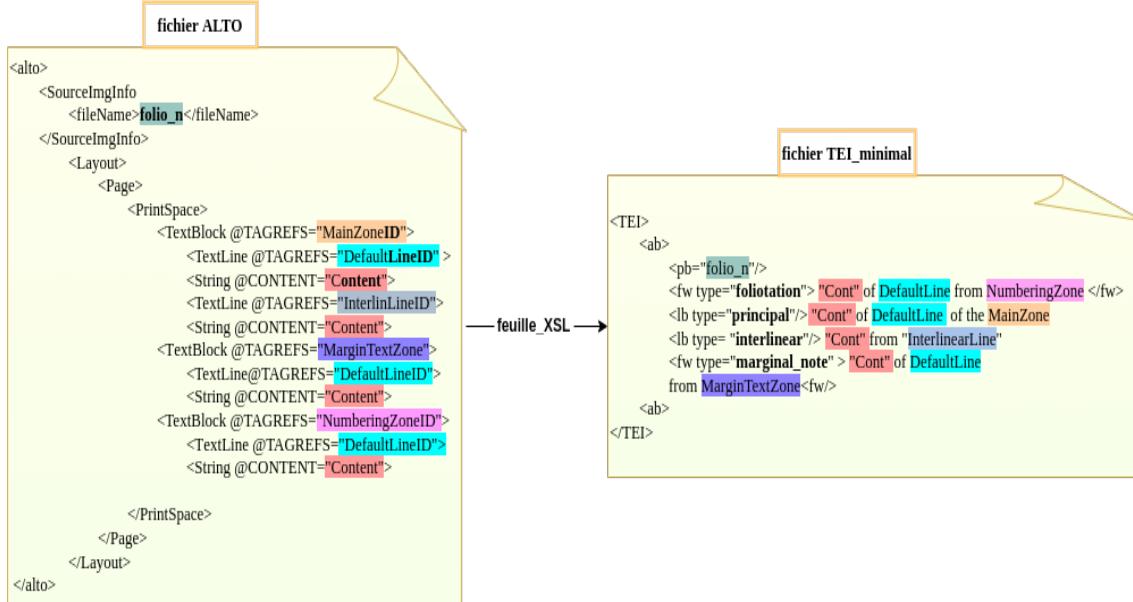


FIGURE 2.3 – Transformation des fichiers ALTO à un fichier XML-TEI minimal avec intégration de l’ontologie SegmOnto

L’interopérabilité et la réutilisation des données, avec une possibilité de transformation en plusieurs format différents, se trouve au cœur des débats des Humanités Numériques et cela pas sans raison. Ainsi, au lieu de tout encoder à la main, une grande partie est structurée de façon automatique, ce qui permet d’économiser du temps tout en assurant la cohérence des données. Il ne reste que compléter l’encodage manuellement, en fixant les éléments d’intérêt. Étant donnée les problématiques énoncées dans la première partie du mémoire, nous n’avons pas envisagé notre encodage comme une simple pratique d’édition, mais plutôt comme une pratique de recherche qui permettra l’interrogation et exploitation des données de manière la plus efficace.

2.2.2 XML-TEI

En se penchant sur la notion de l’édition critique, étant donné la distance entre l’original (vi^e siècle) et les copies (duviii^e au ix^e siècle), ce projet n’envisage ni de remonter à l’original ni de retrouver l’intention de l’auteur et la variante parfaite²⁵. Effectivement, la méthod(ologi)e lachmannienne a véhiculé l’idéologie de la reconstruction de l’*Urtext* archétypique perdu, alors que tout texte connaît des réalisations différentes au cours de son histoire. Cela vaut particulièrement pour les éditions des gloses et des manuscrits glosés, où on considère que

l’ensemble des interventions faites par les acteurs qui l’ont produit constitue la recension primaire de la collection, quel que soit le nombre de strates qui la compose. [...] Dans le cadre de l’analyse des gloses décrit en termes de relations entre collections et corpus, le système stemmatique de Lachmann ne s’applique pas : il serait vain de

²⁵. Bernard Cerquiglini, « Eloge de la variante : Une Histoire critique de la philologie (Paris, 1988) », *Vincent Kauffmann, L’Equivoque épistolaire (Paris, 1990)* ()

vouloir faire remonter une collection de gloses à un modèle identique antérieur, puisqu'il n'a probablement jamais existé²⁶. La démarche stemmatique [...] n'a aucun sens dans un contexte de transmission fluide. Il n'y a pas de texte « original » à reconstruire, sinon une source « ultime », modifiée et parfois transformée, de proche en proche, au fil des réemplois.²⁷.

Des méthodes et des outils d'analyse des relations entre collections et corpus de gloses qui offrent une plus grande souplesse se mettent à la disposition des éditeurs comme la phylogénétique et l'analyse des réseaux²⁸. Le présent projet vise simplement à mettre le texte dans son contexte²⁹ et à explorer les informations textuelles, paratextuelles et métatextuelles offertes par les témoins. Décontextualiser soit le texte de son paratexte soit l'inverse serait, pour citer Vivien Law « rather like the fossil hunter who ignores the strata in which his prize specimen is embedded³⁰ ».

Sur les fondements théoriques des éditions critiques numériques, Peter Robinson³¹ nomme six aspects essentiels à tenir en compte :

1. Une édition numérique critique est ancrée dans une analyse historique des matériaux ;
2. Une édition numérique critique présente des hypothèses sur la création et l'évolution dans le temps ;
3. Une édition numérique critique fournit un enregistrement et une classification dans le temps, dans de nombreuses dimensions et avec des détails appropriés ;
4. une édition numérique critique peut présenter un texte édité, parmi tous les textes qu'elle propose ;
5. une édition numérique critique offre aux lecteurs l'espace et les outils pour qu'ils puissent développer leurs propres hypothèses et modes de lecture ;
6. une édition numérique critique doit offrir tout cela de manière à enrichir la lecture.

Tous ces critères peuvent être satisfaits dans un environnement numérique, tout en se souciant des spécificités du document à encoder. En effet, l'absence de contrainte spatiale permet de représenter les données dans leur totalité et la souplesse des outils numériques rend ces données permis une interrogation plus efficace. Quoi qu'il en soit, pour que l'ensemble de l'information présente dans un manuscrit glosé soit exploitable, sa typologie précieuse doit être reproduite via l'encodage. Ce dernier doit fournir les moyens d'analyser l'ensemble des manifestations selon les

26. La même constatation fait Emmanuelle Kuhry, éditrice du *De plantis* glosé Emmanuelle Kuhry, « Medieval Glosses as a Test Subject for the Building of Tools for Digital Critical Editions », *Journal of the Text Encoding Initiative*–13 (2020)

27. F. Cinato, *Priscien glosé...*, p.260

28. cf. l'édition de E. Steinová et P. Boot, « The glosses to the first book of the Etymologiae of Isidore of Seville : a digital scholarly edition »...

29. Pour une discussion sur les questions numériques que cette recontextualisation implique voir. E. Pierazzo et P. A. Stokes, « Putting the text back into context : a codicological approach to manuscript transcription »...

30. V. Law, *Grammar and grammarians in the Early Middle Ages...*, p.22

31. Peter Robinson et HTM Van Vliet, « What is a critical digital edition », *Variants*, 1 (2001), p. 43-62

différents niveaux d'informations qu'une typologie en « multiple facettes » impose. Les questions essentielles sont liées à la structure, la sélection et l'encodage des données constituant le corpus. Ce corpus de gloses ouvrira autant de possibilités d'exploitation que le contenu des métadonnées sera organisé pour tenir compte des spécificités inhérentes aux gloses. La qualité des résultats et les possibilités offertes par le corpus dépendront de la richesse et de la variété des informations encodées : elles constitueront à la fois autant de critères de sériation que d'outils d'analyse. Dans l'ère du numérique, il n'est plus question d'encoder le plus grand nombre d'informations, mais plutôt de faire du document « une synthèse structurée des variétés observables ». Le contenu fondamental des collections - lemmes et gloses - devra être encodé de manière à permettre l'application de technologies visant à faciliter l'exploitation et l'exploration scientifiques des données.

2.2.3 Choix de l'encodage

Dans cette section on va esquisser des stratégies de modélisation de l'encodage qui conviennent d'avantage aux manuscrits grammaticaux glosés, et qui remplissent au mieux les exigences d'un matériel et une typologie complexes, voire constituent une condition *sine qua non*. Déjà les avantages d'une approche numérique concernant les manuscrits glosés ont été évoqués par Franck Cinato³² et Paolo Monella³³, qui ont travaillé et continuent à le faire sur la modélisation et l'encodage de la tradition manuscrite de Priscien de Césarée, et dont le travail a servi de point d'appui à notre processus d'encodage proprement dit. S'aligner sur les principales pratiques du domaine est crucial pour l'interopérabilité et l'homogénéité des données qui s'apprêtent à un examen vis à vis, comme pour le grand réseau de corpus des *Grammatici Latini*³⁴.

L'encodage doit fournir les moyens d'analyser l'ensemble des manifestations selon les différents niveaux d'informations dégagés précédemment. Face à ces niveaux, Franck Cinato a adapté des concepts développées dans le cadre plus large de l'organisation des connaissances sur le cas des manuscrits glosés, en suivant le *faceted classification system*³⁵. Ces facettes permettent une caractérisation complète de l'entité qu'on appelle glose, et consiste en plusieurs dimensions, entre autres son emplacement (interlinéaire ou marginale) sa forme (signe, morphème, syntagme, phrase), son sens (classification qualitative selon l'apport d'une glose) et l'acteur (la main)³⁶. L'intérêt qu'une typologie complexe pendant la sériation permettra de réaliser réside en la « finesse granulométrique » dans les opérations de tri appliquées à de grands ensembles de données. Un encodage qui inclut

32. F. Cinato, *Priscien glosé...* et F. Cinato et E. PARIS, « Perspectives offertes par un corpus électronique de gloses sur Priscien »...

33. P. Monella, « A digital critical edition model for Priscian »...

34. Sur ce sujet la vision et l'entreprise majeure de l'Index Grammaticus déjà pendant les années '90 : V. Lomanto et Nino Marinone, *Index grammaticus : an index to the Latin grammar texts*, t. 81, 1990 et les perspectives d'un corpus électronique des GL décrites par Prof. Alessandro Garcea dans Alessandro Garcea, Clément Plancq et F. Cinato, « Corpus Grammaticorum Latinorum : un projet de traitement informatique autour des grammairiens latins », *Corpus grammaticorum latinorum* (, 2010), p. 1000-1024

35. Une classification à facettes utilise des catégories sémantiques, soit générales, soit spécifiques à un sujet, qui se combinent pour créer la base de classification complète. Des facettes subordonnées affinent davantage le sujet.

36. Pour l'éventail complet voir. F. Cinato, *Priscien glosé...*, Annexe 3

l'ensemble des informations pivot, comme il sera mis en évidence lors de la visualisation des données (cf. *infra*), qui s'avère particulièrement utile pour observer les mouvements de la confection du manuscrit et les préoccupations des glossateurs-compilateurs.

Le guide complet de l'encodage du texte se trouve dans le fichier ODD, disponible en version PDF et XHTML dans notre dépôt [GitHub](#). Nous nous contenterons ici de justifier la gestion de l'information multi-facette dans le choix de l'encodage.

MISE EN PAGE

En considérant que la mise en page joue un rôle fonctionnel plutôt qu'accessoire dans la disposition du contenu intellectuel de l'oeuvre, il est impératif que l'encodage inclue ces spécificités. En effet, la description « codicologique » qui s'attache plus à l'aspect formel des gloses, constitue une première facette du système d'analyse. Deux aspects qui méritent notre attention sont 1) la disposition en colonnes qui rompe le sens de lecture de l'horizontale à la verticale et 2) le type des lignes selon leur environnement et leur fonctionnalité.

En ce qui concerne la typologie des lignes, l'encodage, produit de la transformation XSL se présente ainsi : Les *DefaultLines* se sont transformées en

`<lb type="principal">`

et les *InterlinearLines* en :

`<lb type="interlinear">`

Une spécification au niveau des lignes permet de la sélection et l'extraction, en notre gré, soit uniquement du texte principal (utile pour la collation du témoin), soit uniquement des interlignes.

LEMMES

Comme il est déjà mentionnée, tout mot ou groupe de mots d'un texte principal forme potentiellement un lemme. Les lemmes, dans un sens ontologique constituent des segments arbitraires du texte principal qui se comportent en tant entités possédant un degré d'autonomie relatif à leur dépendance du texte. Pour en fournir des exemples spécifiques, un verbe isolé (par exemple *amo*) peut se suffire à lui-même, indépendamment du contexte immédiat ou intellectuel de la phrase. Par contre, un pronom (par exemple *a quibus*) dépend entièrement de son contexte pour faire sens. Selon la définition du lemme adopté tout segment dont le sens a été augmenté en n'importe quelle manière, constitue un lemme et est unique.

Côté pratique, un lemme est encodé ainsi :

<seg type="lemma" xml:id="f02r_101.2">nouas qūestiones</seg> :

- un élément neutre **seg** qui évite tout sémantisme en soi ;
- un attribut **type** = "lemma" qui spécifie la valeur sémantique du segment ;
- un **xml :id** composé des coordonnées du segment : ici il s'agit du deuxième lemme de la première ligne du folio 2r.

GLOSES IN-SITU ET GLOSES MARGINALES

En général, les gloses interlinéaires peuvent être non verbales, c'est-à-dire des signes de construction, des marques prosodiques, des chiffres (romains et plus tard également arabes), des petites lettres et d'autres symboles, ainsi que verbales, expliquant divers aspects introductifs mais essentiels du texte. Dans le domaine de la prosodie, les gloses traitent généralement des questions d'accents, de mètre et de techniques poétiques. Les questions de grammaire, tant morphologiques que syntaxiques, sont illustrés par la compléTION de prépositions omises, l'explication des cas, la clarification des sujets et des objets, surtout lorsqu'ils sont exprimés par des pronoms (assez souvent dans VLO41), et l'aide concernant l'ordre des mots et la subordination des clauses. En matière de vocabulaire, les glossateurs proposent normalement des synonymes (rarement des antonymes, parfois la négation des antonymes), donnent des équivalents latins pour les mots grecs ou des traductions vernaculaires pour les mots latins, et fournissent des noms qui aident et enrichissent la compréhension du lecteur. En raison de leur nature très localisée, ces types de gloses sont généralement placés entre les lignes et à proximité de leurs lemmes, mais en raison de limitations particulières, ils peuvent être déplacés, parfois même jusqu'aux marges (cf. le premier critère du guide d'attribution des mains [ici](#)). Dans de tels cas, le jugement critique du lecteur est requis afin de les relier à la place qui leur est destinée dans le texte. Du point de vue de l'éditeur, toutes ces informations sont importantes et donnent un aperçu des pratiques pédagogiques qui découlent de la pratique des gloses.

Une typologie formelle des gloses attestées dans les manuscrits grammaticaux a déjà été formulée par Rijcklof H.F. Hofman³⁷ à la fin des années '90, typologie augmentée et épurée par Franck Cinato pour son édition de Priscien glosé. Cette facette « Sens » consiste en sept souscatégories principales, à savoir :

- S₁ Prosodique
- S₂ Lexicale
- S₃ Grammaticale (morphologique)
- S₄ Syntaxique
- S₅ Explicative (commentaire)

³⁷. Hofman, Rijcklof HF. The Sankt Gall Priscian commentary. Part 1. 2. Translation and commentary ; Indices. Nodus-Publ., 1996.

- S6 Ecdotiques
- S7 Notes socio-historiques

qui acceptent toutes des augmentations de sorte qu'il existe un chiffre unique pour toute manifestation sémantique de glose³⁸. Par exemple la typologie de loin la plus fréquente dans le VLO41 est la *S₂₂*, la glose lexicale qui procure un simple synonyme ou la *S₂₃* qui attribue une définition, ou bien la *S₅₄* qui propose une étymologie.

En même temps, bien qu'il faille prendre toutes les précautions nécessaires pour analyser ces données, il apparaît essentiel, pour l'établissement de la diachronie au sein des collections, de signaler les changements de scripteurs³⁹. Il convient d'abord de distinguer mains de copistes et de glossateurs. Par convention, on attribuera un chiffre aux copistes (mains 1, 2, 3, etc.) et une lettre aux annotateurs (mains A, B, C, etc.). En fonction du type de collection, les éléments du groupe lexical (lemme et glose) n'émaneront pas des mêmes personnages.

En pratique cette modélisation théorique donne l'encodage suivant :

```

<lb type="principal"/> [...]gur auguro as, in as ut <seg type="lemma" xml:id="f11v_122_a">
uan </seg>
<gloss type="S22" xml:id="f11v_122_a" target="#f11v_122" resp="#B">
uan </gloss>
<gloss rend="italic" type="S23" xml:id="f11v_122_b" target="#f11v_122" resp="#A">
recedens a lege</gloss>
<gloss type="S23" xml:id="f11v_122_c" target="#f11v_122" resp="#D">
‡ inutilis</gloss>

```

L'élément déjà défini⁴⁰ par la TEI *gloss* est utilisé pour encadrer les gloses, suivi par l'attribut *type* qui sert à préciser la typologie mis en avant par Cinato. Pour ce qui est de plus, les gloses possèdent leur propre **xml :id** qui consiste en celui du lemme correspondant avec l'extension *_a* ou *_b* ou même *_c* selon l'ordre d'apparition au cas de plusieurs gloses afferant à un et le même lemme. Finalement, la main responsable est désigné par l'attribut *resp*.

Pour le contenu des gloses qui se trouvent en marge, annotées en tant que MarginalTextZone avec SegmOnto, et qui s'étalent souvent en plusieurs lignes, sont encodés ainsi :

38. Pour la typologie complète on renvoie à l'Annexe 3 de *Ibid*.

39. Un petit guide des critères d'attribution des mains avec description détaillée et un échantillon se trouve dans le document *hand_attribution_guide* de notre dépôt GitHub.

40. Selon les Guidelines de la TEI : <gloss> (gloss) identifies a phrase or word used to provide a gloss or definition for some other word or phrase.

```

<fw type="marginal_note" xml:id="f07r_124.1_b" corresp="#f07r_124.1" resp="#A">
    <lb type="principal"/> .s. nubo nubilis
    <lb type="principal"/> nubilis dř apta ad
    <lb type="principal"/> muliem duxi
</fw>

```

L'élément `<fw>` (forme work), qui contient un titre courant (par exemple, un en-tête, un pied de page), un mot d'accroche ou un élément similaire apparaissant sur la page actuelle, sert comme élément qui encadre la note marginale. De nouveau, suivant l'exemple des gloses interlinéaires, des `xml:id` y sont attribués et la main responsable est indiquée.

GESTION DES COLONNES

Un encodage propre aux colonnes, élément si important pour la transmission du texte, est adopté afin de sauvegarder l'ordre des lignes, ainsi que les couples lemmes-gloses présents en leur sein. L'exemple suivant est un extrait d'encodage de deux colonnes successives du folio 4r (il en existe au total 3), dont certaines lignes principales sont glosées. De nouveau on se sert de l'élément `<fw>`, cette fois de type "colonnes" pour encadrer les groupes de colonnes. Au sein de cette structure, des éléments vides `<cb>` numérotés, indiquent l'ordre d'apparition dans la page.

```

<fw type="colonnes" n="f04r">
    <cb n="1">
        <lb n="01" type="principal"/> ut
        <seg type="lemma" xml:id="f04r_c1_101" >nuntius</seg>
        <lb type="interlinear"/>
        <gloss type="S22" xml:id="f04r_c1_101_a" target="#f04r_c1_101" resp="#B">
            <lb n="02" type="principal"/> nuntio. nutias.
        <lb n="03" type="principal"/>
        <seg type="lemma" xml:id="f04r_c1_103" >Sotius</seg>
        <lb type="interlinear"/>
        <gloss type="S22" xml:id="f04r_c1_103_a" target="#f04r_c1_103" resp="#B">
            <lb n="04" type="principal"/> Sotio. sotias
        <lb n="05" type="principal"/>

```

```

<seg type="lemma" xml:id="f04r_c1_105">Sautius</seg>

<lb type="interlinear"/>
<gloss type="S22" xml:id="f04r_c1_105_a" target="#f04r_c1_105" resp="#

[...]

<cb n="2"/>

<lb n="01" type="principal"/> Radio. radias.

<lb n="02" type="principal"/> consilium

<lb n="03" type="principal"/> consilioī ēsiliaris

<lb n="04" type="principal"/>
<seg type="lemma" xml:id="f04r_c2_104">conciliū </seg>

<lb type="interlinear"/>

<gloss type="S22" xml:id="f04r_c2_104_a" target="#f04r_c2_104" resp="#B">

[...]

</fw>

```

CITATIONS ET *GRAECA*

Au couple indissociable lemme-glosse et à la mise en page s'ajoutent deux aspects clés du *De uerbo* et de sa tradition textuelle : les *graeca* et les citations. Ces deux éléments, comme il a été déjà évoqué, ont une signification philologique et culturelle particulière pour l'histoire des textes linguistiques.

Comme mentionné ci-dessus, le *De uerbo* a été composé à Constantinople, c'est-à-dire dans un environnement de langue grecque destiné à un disciple hellénophone, ce qui explique pourquoi Eutyches, cependant pas aussi souvent que son maître Priscien, compare les conjugaisons des verbes latins et grecs, et semble parfois traduire en grec afin d'expliquer le latin, principalement dans le deuxième livre. Si la présence des équivalents grecs répondait aux besoins du bilinguisme grec-latine de la Constantinople du VIe siècle, elle a ensuite posé un énorme problème aux scribes de l'Europe latine médiévale, qui ont progressivement perdu la connaissance du grec. Effectivement, si ils ne sont pas omis tout court, les *Graeca* portent des erreurs significatives, précieuses pour la *recensio*.

Dans l'exemple suivant tiré du folio 23v, Eutychès ressent le besoin de se référer au terme grec pour "fond", à savoir ΠΤΘ Μ Η Ν, que le copiste a retranscrit avec l'orthographe phonétique, précieux indice pour la prononciation des termes grecs.

[...] a nomine qđ est fundus id est <**foreign** **xml:lang="grc"**> ΒΤΘΕΜΕΝ.</**foreign**>

D'autre part les citations, tirées d'œuvres littéraires, sont également précieuses pour la *recensio*. Ces dernières offrent une piste stable contre laquelle la « Latinité » du copiste ou de l'exemplar peut être mise en examen. Une petite parenthèse concernant les citations des auteurs classiques et leur importance pour la transmission des savoirs pendant le Haut Moyen Âge : il suffit de mentionner que de toute la littérature profane, les manuscrits contenant les *Artes* de l'Antiquité tardive sont les seuls qui soient parvenus à les conserver, de siècle en siècle. Même si on avait cessé de recopier Cicerón, Virgile, Horace, en continuant à recopier Donat, Charisius, Phocas, Priscien et les autres grammairiens de l'Antiquité tardive, que l'on a pu un jour reprendre goût à, lire et à recopier Virgile, Cicerón, Horace. Cette circonstance seule suffirait à établir que les manuscrits grammaticaux ne sont guère des recueils comme les autres⁴¹. Luca Martorelli, a dressé, en enrichissant les références de Lindemann⁴² une liste détaillée des 141 citations qui figurent chez Eutychès accompagnée des notes critiques sur les endroits où Eutychès d'écarte de l'original⁴³. C'est à l'aide de cette liste que nous avons pu localiser avec précision les nombreuses citations et leur associer un lien vers l'édition critique la plus récente, à condition que ça soit en ligne⁴⁴. Exemple d'encodage :

```

<lb type="principal" break="no"/> VIII
<quote type="poetry"> has forte inuisitis aras
  <ref target="http://www.perseus.tufts.edu/hopper/text?doc=
    Perseus\%3Atext\%3A2008.01.0498\%3Abook\%3D1" resp="#LM">
    Stat. Theb. 1, 668</ref>
</quote>

```

Dans un environnement **<quote>**, on indique le genre auquel appartient la citation, prose ou poésie, ainsi que le responsable (@resp⁴⁵) des coordonnées de la référence (à l'intérieur d'un **<ref>**) et le lien (à l'aide de @target) vers le passage original.

Une granularité fine, adaptée à nos données, permet l'exploitation du texte à deux volets. En premier lieu, préserver l'information de la mise en page, et en deuxième lieu, mieux interroger le texte en extrayant les balises d'intérêt pour une comparaison interne ou entre témoins.

41. Louis Holtz, « La typologie des manuscrits grammaticaux latins »..., p.252

42. Friedrich Lindemann, *Corpus grammaticorum latinorum veterum*, t. 3, 1833

43. Luca Martorelli, « Le citazioni in Eutiche », *Revue de philologie, de littérature et d'histoire anciennes*, 91-2 (2017), p. 55-88

44. notamment vers le [Perseus Digital Library](http://www.perseus.tufts.edu/hopper/text?doc=Perseus\%3Atext\%3A2008.01.0498\%3Abook\%3D1), ou bien sur archive.org.

45. Dans la plupart des cas, il s'agit de Luca Martorelli, sauf lorsque j'ai dû faire référence à une édition différente, disponible en ligne (et changer par la suite les coordonnées de référence). Dans ce cas la @resp est moi.

2.2.4 Questions de normalisation/lemmatisation

Bien que la TEI offre la possibilité de normaliser le texte tout en conservant les leçons originales, nous avons décidé de ne pas procéder à une normalisation. Ceci est dû à la nature particulière du contenu grammatical et à la façon dont nous voulons traiter leurs propos. Et s'il est important d'utiliser les outils offerts par le numérique, tous les outils ne sont pas forcément adaptés à tout type de données. On est pas les premiers à être confronté à cette question. Déjà, dès les années '90, en exposant la méthodologie pour la confection de l'*Index Grammaticus*⁴⁶, Valeria Lomanto et Nino Marinone se déemandent quelle est la meilleure pratique à adopter pour étudier les co-occurrences des termes dans le corpus des *GL*. En considérant l'option de créer un *Lexicon* (un thesaurus constitué des entrées d'un dictionnaire), ils remarquent que :

[...] it is not very *suitable* for grammarians in that it involves *lemmatization* and, as it provides *no context*, it cannot reveal one of the salient characteristics of the *artes*, namely its use of a sort of formulaic language present with *slight variations* in every work. [...] The decision not to lemmatize derives both from the particular nature of the texts and from the future users of the concordance. In the *artes* a high percentage of forms is represented by *isolated graphemes* or by isolated morphemes or, in the chapters on metre, by units which are not linguistic but rhythmic (e.g. *armaui rumqueca* etc.). Moreover, lemmatization, while it requires a reliable critical text, necessarily directs, if it does not condition, the interpretation.

Une intervention éditoriale qui implique une corréction en masse des « erreurs » que les copistes commettent, comme c'est le cas de la démarche de Keil, va à l'encontre d'une étude scrupuleuse des variantes grammaticales utilisées par les copistes, qui donne des amples informations sur leur propre niveau de latinité. Pour reprendre les mots de Vivien Law⁴⁷ :

It is of the utmost importance for future research that editors be scrupulous in their treatment of the texts before them. All departures from the text offered by the manuscripts should be signalled. Emendations, particularly in passages borrowed from Classical grammarians, should be undertaken sparsely and indicated as such in the text. Editorial policy, with regard to orthography should be stated at the outset : the reader should be aware of how far the text before him differs from what a medieval student might have read. There is no justification when editing a medieval grammatical text for wholesale 'normalisation' to Classical conventions ; such a practice will cause havoc in alphabetical lists and reflect a misleading image of the author's latinity.

Une normalisation consiste à réduire les majuscules des noms communs, à uniformiser les orthographes multiples qui apparaissent pour un même terme sémantique, les dates et les chiffres et à développer les abréviations, pour procéder à une lemmatisation. La lemmatisation associe à

46. V. Lomanto, « A concordance to Keil's Latin grammarians », *Computers and the Humanities*, 24-5 (1990), p. 427-435

47. V. Law, *The insular Latin grammarians*, 1982, p.107

ces graphies normalisées un lemme correspondant à l'entrée d'un dictionnaire et une catégorie grammaticale. Par conséquent, elle met au premier plan le niveau sémantique, puisque sa fonction principale consiste à définir les différents sens de chaque terme, à établir des relations de sens et à tracer la frontière poreuse entre polysémie et homonymie. Ce serait donc l'aboutissement d'une tradition philologique et exégétique si solidement ancrée, à l'issue d'une édition, qu'elle permet au savant d'affronter les problèmes de sens les plus délicats.

Pour ces raisons, nous avons opté pour une édition documentaire qui reflète au mieux l'état de la langue présenté par le manuscrit. Les éditions dites diplomatiques des grammaires pourraient constituer une ressource particulièrement précieuse pour les historiens de l'éducation et de la linguistique, car les enseignants et les penseurs médiévaux ne disposaient pas d'éditions critiques idéalisées dans leurs bibliothèques. Ils devaient se contenter des copies réelles à leur disposition, avec toutes les erreurs, les malentendus et les omissions qu'elles transmettaient.

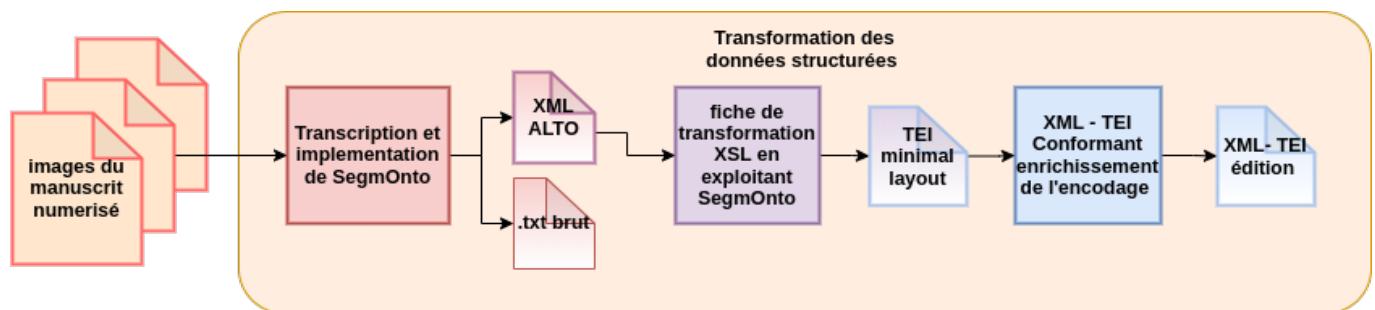


FIGURE 2.4 – Rappel du pipeline : Modelisation des opérations de transformation des données brutes en données structurées jusqu'à l'encodage complet (axe horizontale)

Après avoir effectué l'édition documentaire de notre témoin, il convient de se lancer dans la troisième étape, à savoir l'analyse exploratoire de nos données à multiples facettes. Un petit *caveat* avant de continuer : Une utilisation plus propice de notre texte encodé en XML-TEI serait l'extraction des balises qui nous intéressent dans un tableau .csv, pour la visualisation des données issues de l'encodage. Pour la première année du Master, nous nous sommes limités à manipuler un fichier .csv externe, compilé indépendamment du fichier XML-TEI et qui nous a servi de point de repère pour son encodage. Pour l'année prochaine, il contribuerait largement à l'homogénéité des données d'exporter les balises d'intérêt pour l'étape de la visualisation.

2.3 Visualisation des données

L'approche typologique des gloses, au-delà d'un simple classement, permet de mettre en évidence les préoccupations des glossateurs, ainsi que la spécificité de leurs enseignements⁴⁸. Il s'agit

48. Toujours suivant : F. Cinato et E. PARIS, « Perspectives offertes par un corpus électronique de gloses sur Priscien »...

des informations « externes », liées à leur transmission, une sorte de métadonnées, aspects d'aillieurs proprement paléographiques, qui se révèlent d'une importance primordiale dans le cadre de l'étude des gloses. La distinction des écritures détermine des éléments essentiels à l'analyse. Les premiers qui viennent à l'esprit ont trait à la distinction des mains qui ont laissé des couches superposées des gloses : un même manuscrit a souvent été étudié en différents lieux et à différentes époques. Il faut donc – pour rendre compte de la perspective diachronique au sein de la collection – établir une chronologie relative des mains qui ont participé à la formation du *peri-texte*. Cette étape n'est possible que dans le cas des sources de type Γ , comme le VLO⁴¹.

Une facette spécifique « Acteur » est prévue *a priori* par Franck Cinato pour les mains différentes, en fonction du rôle de chaque glosateur, d'une échelle allant du 1 à 3 où :

- A₁ : Glossateur-copiste
- A₂ : Glossateur-compilateur
- A₃ : Glossateur-exégète

Faute d'expérience de notre part et en se fondant sur un échantillon encore limité de données, nous avons décidé de ne pas procéder à une *a priori* attribution de rôles. En revanche, à l'aide des outils numériques, nous avons décidé de tenter une « analyse exploratoire » sur les statistiques fournies par l'attribution des mains, aux fins de déterminer quelques caractéristiques propres à chaque glosateur. En effet, les outils numériques sont particulièrement adaptés pour la manipulation à large échelle et pour la visualisation des données à multiples facettes (comme le réseaux des gloses au sein du VLO⁴¹).

Ainsi grâce aux scripts python⁴⁹ et la librairie *matplotlib*, en choisissant des graphiques adaptés à nos données multi-variables, on a tâche à faire des observations générales sur les mouvements successifs d'annotation du manuscrit et sur les préoccupations différentes des glossateurs-compilateurs.

2.3.1 Mouvements d'annotation

Pour la première visualisation, nous nous sommes intéressés aux variables suivantes : la quantité de gloses écrites par et par folio, ce qui donne deux graphiques complémentaires (Figure 2.5 a et b).

49. Disponibles dans les notebooks de notre dépôt <https://github.com/malamatenia/Eutyches/tree/main/python-tools>

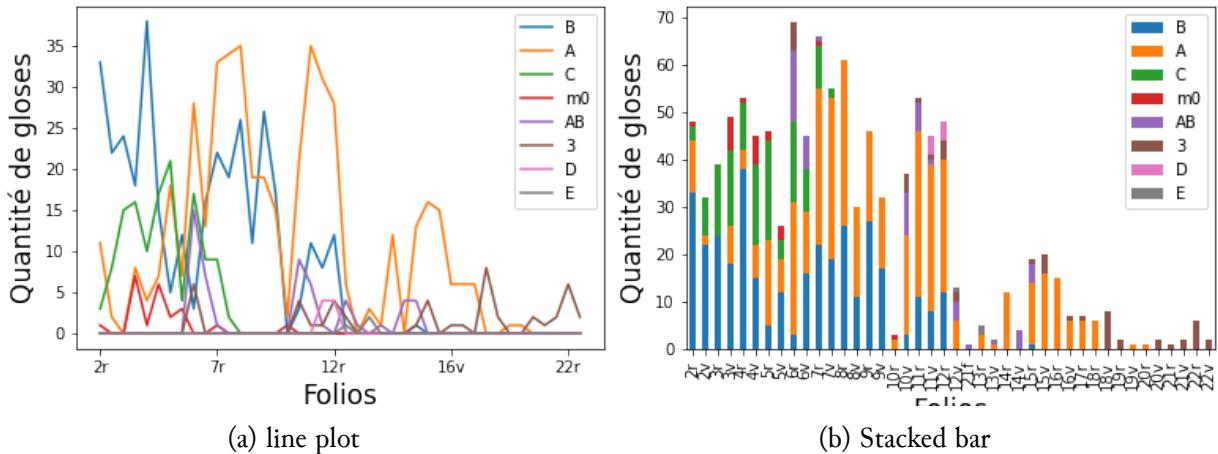


FIGURE 2.5 – Dans le premier graphique, une ligne correspond à une main en traçant sa contribution per folio. Le deuxième prend les mêmes variables et en donne la densité de la contribution de chaque main.

Mg	
3	48
A	456
AB	53
B	338
C	130
D	8
E	3
m0	22

Name: quantity, dtype: int64

FIGURE 2.6 – Somme des gloses écrites par chaque main. "Mg" vaut pour Main de glosateur.

Le premier graphique permet de suivre le mouvement de chaque main par folios, l'intensité de leur activité au fur et à mesure. La seconde, complémentaire de la première, permet d'examiner plus en profondeur le pourcentage de contribution des mandants par folio, et l'émergence de tendances. Des observations d'ordre générale peuvent être faites :

- La quantité des gloses se diminue au fur et à mesure de la progression des folios, en confirmant la tendance générale qui atteste la concentration des gloses au début des manuscrits ;
- Le rôle marginal des 4 dernières mains est vraiment accentué par la visualisation. La main B est la main dominante pour les 5 premiers folios⁵⁰, puis elle coexiste avec la main A jusqu'au folio 10v et la main A devient la main dominante jusqu'à la fin ;
- Trois mains apparaissent comme les principaux glossateurs du VLO41, à savoir A, B, et C. Alors que les deux premières ont une répartition plus équilibrée sur l'ensemble de la surface glosée, A prenant le relais après B, la troisième, C, se concentre uniquement dans les 14 premiers folios.

⁵⁰. Nous précisons ici que nous avons nommé les mains selon l'ordre d'apparition dans le manuscrit et non selon l'importance ou la quantité de gloses.

- On constate une absence générale de gloses dans le folio 10r.

Les observations quantitatives nous servent uniquement d'outil d'analyse et ne doivent pas être considérés comme une source suffisante d'interprétation. Néanmoins combinées avec les observations qualitatives, elles peuvent être nuancées et ainsi contribuent à des interpretations intéressantes, notamment quand il s'agit des corpus hétérogènes et multivariables.

Les deux dernières constatations méritent une élaboration plus ample. En ce qui concerne l'activité de la main C, il semble qu'elle ait utilisé le manuscrit principalement à des fins pédagogiques, en prenant les premières pages du traité en tant que matériel du cours. Cette interprétation est soutenue par les *notae*⁵¹ présentes sous sa plume. 2v, 6r et 6v et s'arrêtent complètement après le folio 7r, a peu près avant la fin de la main C.

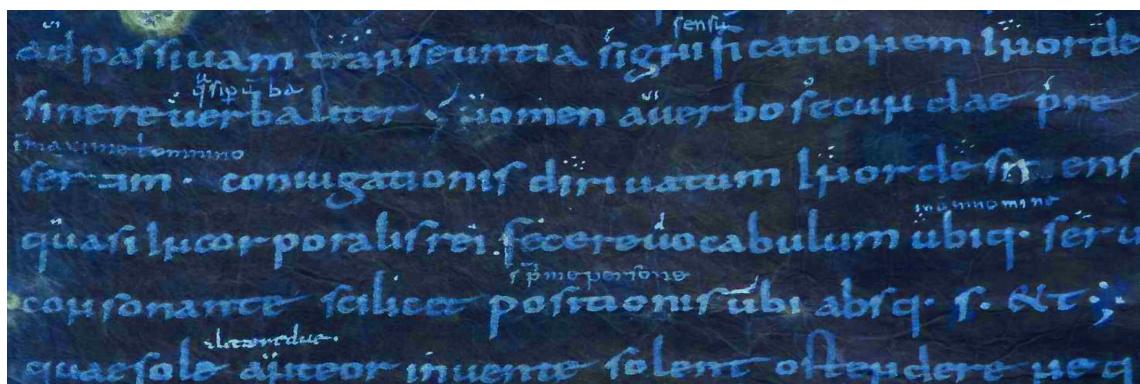


FIGURE 2.7 – Negatif du folio 6v avec échantillon des *notae* de la main C

- Ligne 2 à la fin : **[s]** pour *scribe* → signe indiquant le début d'un extrait ou d'une leçon, dans ce cas en combinaison avec un autre signe indiquant la fin. Une alternative pourrait être *lege*.
- Ligne 4 au milieu : **[d]** pour *dimitte* → signe indiquant la fin d'un extrait ou d'une leçon ; dans ce cas, en combinaison avec un autre signe indiquant le début.
- Ligne 3 au milieu : variation du *trigon* → Quant au trigon et à ses variations comme le quadrilatère, il s'agit d'un caractère assez divers, qui sert souvent comme signe d'attention lors de la lecture.
- Ligne 4 au début : **[n]** sur *quasi* → En l'absence d'une explication concrète de ce que « n » pourrait signifier, nous ne souhaitons pas forcer une interprétation.
- Ligne 1 au début : le chiffre **[VI]** ; En général, les chiffres ou les premières lettres de l'alphabet sont utilisés comme signes de l'ordre syntaxique au sein de la proposition. On observe néanmoins que le chiffre VI dans ce cas est utilisé pour marquer notamment les prépositions⁵². En

51. Les informations sur les signes viennent du : E. Steinová, *Notam superponere studui : The use of technical signs in the early Middle Ages*, thèse de doct., Utrecht University, 2016

52. Qu'au total ce marqueur apparaît 11 fois : 6 fois sur des prépositions (folio 2v in, in, ex, folio 6r ad, a, in) 1 fois sur l'adverbe sic, une fois sur la conjonction si, une fois sur *prius*, une fois sur *unius* et deux fois sur des noms, à savoir sur *uerborum* (folio 2v) et *deriuacione* (folios 5v). Une utilisation stable est difficile à trancher.

effet, Selon le classement des *partes orationis* qui n'est pas complètement fixe pendant l'Antiquité Tardive, la préposition obtient soit la cinquième (Priscien), soit la septième place (Donat) toujours entre l'interjection, l'adverbe ou la conjonction, à savoir les *partes indeclinabiles* dont la position est une *quaestio en soi*⁵³. Par contre, selon Bernard Colombat⁵⁴ ce classement commence à changer pendant le XVIe siècle, Alde Manucius et Gérard Vossius⁵⁵ plaçant la préposition en sixième place. Une corrélation directe entre ces observations est néanmoins hasardeuse, dans le sens où cela impliquerait une date beaucoup plus tardive des gloses écrites par C que celle supposée par DeMeyer. Par conséquent, sans avoir fait un examen approfondi des autres témoins et de la théorie de la *partes orationis* pendant le Moyen-Âge, nous nous limitons à signaler cette corrélation.

Pour ce qui est de l'absence générale de gloses dans le folio 10r, cette coupure coïncide avec le changement de cahiers dans la composition du manuscrit. Entre les folios 9v et 10r, une feuille est visiblement déchirée, en réduisant le quinon initial en quaternion. Nous pouvons supposer soit que cette partie du texte n'a pas porté autant de gloses que les passages avoisinants, soit que le déchirement d'un feuillet, une intervention codicologique externe a eu un impact sur les pratiques d'annotation du manuscrit. Une fois que plusieurs témoins auront été transcrits et comparés, une image plus claire pourra se dessiner sur cette particularité.

2.3.2 Typologies générales vs. spécifiques

Il convient maintenant de se pencher sur la facette suivante. En ce qui concerne la typologie des gloses, nous avons cherché à fournir d'abord un panorama général de la distribution de cette typologie par main, qui donne le graphique de la Figure 2.7. Nous avons choisi un diagramme à barres circulaire par ordre croissant parce que tout autre type de graphique, à cause de l'hétérogénéité des variables, tend à suraccentuer les valeurs élevées, soit des mains, soit du type de glose.

⁵³. Diom. 1.300.26 *partes orationis sunt octo, nomen pronomen verbum participium adverbium coniunctio praepositio interiectio* (d'autres définitions dans : Char. 193.7 ; Dosith. 7.389.9 ; Prob. 4.51.18 ; Don. 585.4, 613.3 ; Cons. 5.338.4 ; ps. Asp. 5.549.19). Samantha Schad, *A lexicon of Latin grammatical terminology*, 2007, s.v. « pars orationis »

⁵⁴. Bernard Colombat, « LES «PARTIES DU DISCOURS»(PARTES ORATIONIS) ET LA RECONSTRUCTION D'UNE SYNTAXE LATINE AU XVI^e SIÈCLE », *Langages*–92 (1988), p. 51–64 et son tableau des classements à la page 5.

⁵⁵. (ce dernier étant un érudit et grammairien, auteur de l'œuvre monumentale *Aristarchus, sive de arte grammatica* du 1635 (cf sa notice détaillée dans http://ctlf.ens-lyon.fr/n_fiche.asp?cod=1257) et d'ailleurs père d'Isaac Vossius, possesseur du manuscrit au XVII^e siècle)

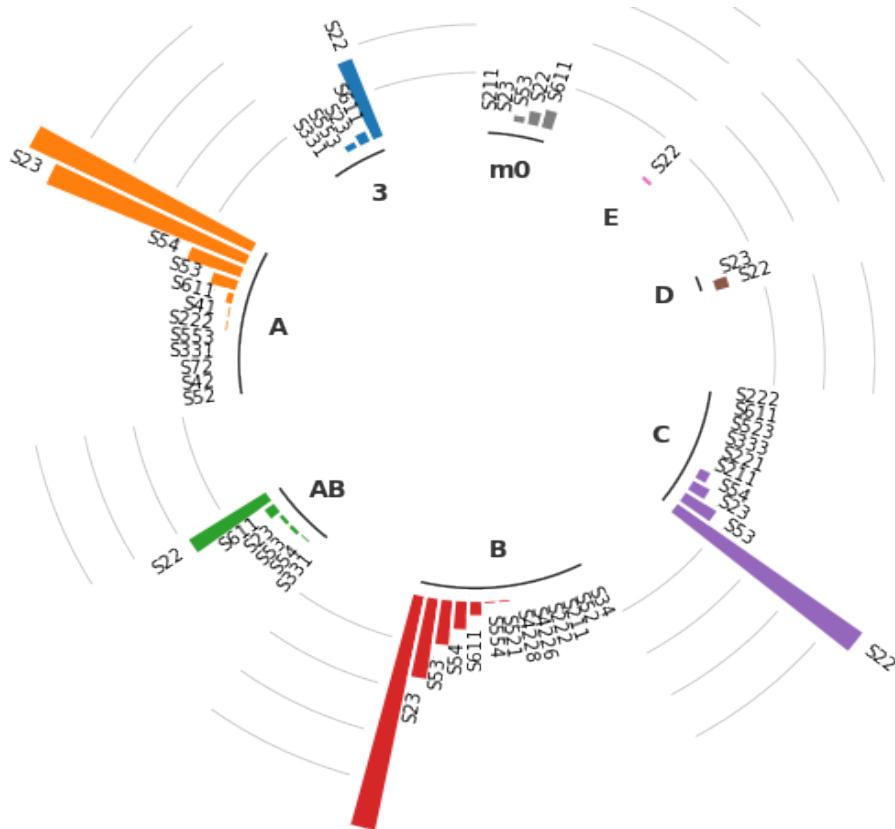
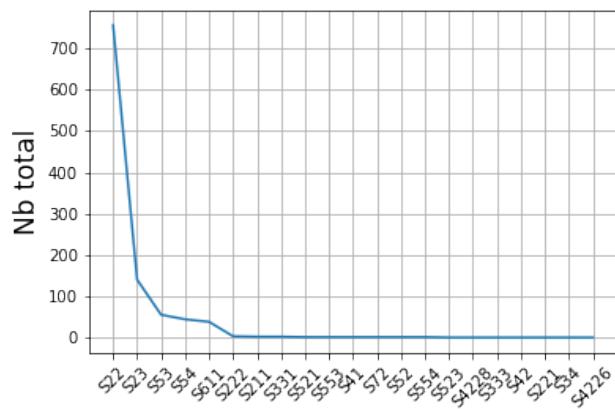


FIGURE 2.8 – Circular barplot pour la distribution hiérarchique du type de gloses utilisé par main.

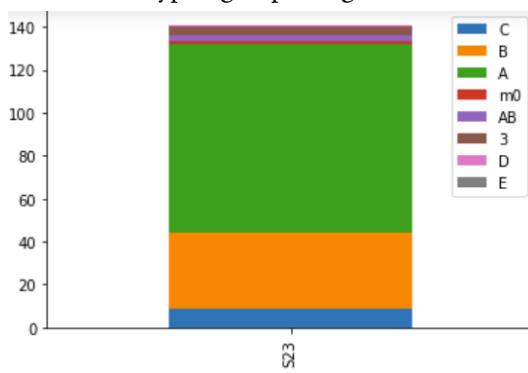
Il devient évident qu'une certaine typologie est privilégiée par toutes les mains, y contribuant à des taux différents, en tenant toujours compte de la quantité de gloses que chaque main a écrites.

- [S22] Synonyme ;
- [S23] Définition ;
- [S53] Précisions sur le texte, glose élucidant le sens ;
- [S54] Étymologie ;
- [S611] Correction critique du texte (y compris des ajouts postérieurs).

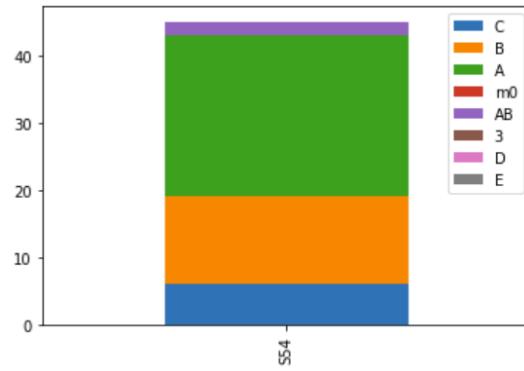
Un tableau complet se trouve ci-dessous :



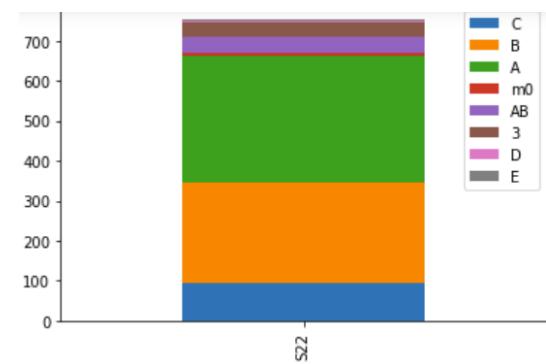
(a) Typologies privilégiées



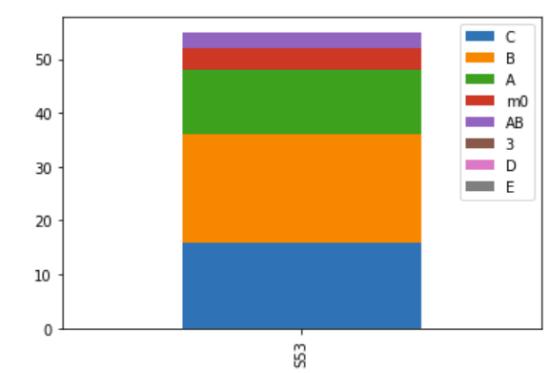
(c) Définition



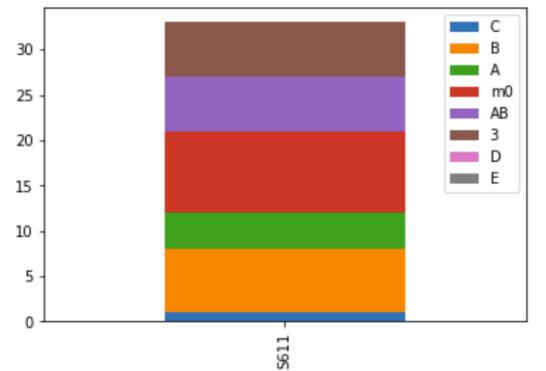
(e) Étymologie



(b) Synonyme



(d) Élucide les ambiguïtés du texte



(f) Intervention éditoriale : corrections

Outre les tendances générales, qui constituent les cas de base lors d'une annotation, ce sont les typologies moins fréquentes qui font la différence dans le caractère d'une campagne d'annotation ou le « style » d'un glosateur. Plus précisément :

Pour A :

- S41 Signe de construction syntaxique (2 occurrences)
- S42 Mot(s) explicitant la construction (1 occurrence)

Pour B :

- S554 Commentaire mythologique (2 occurrences)
- S521 « quia » gloses (2 occurrences)
- S34 Indication d'adverbe (1 occurrence)
- S4226 supplétive par un adverbe (1 occurrence)
- S4228 supplétive par un verbe (1 occurrence)

Pour C :

- S333 morphologie du *modus* (2 occurrences)
- S221 négation d'un antonyme (1 occurrence)

Si peu nombreuses que ce soient ces occurrences, les groupes concernés, grâce à leur spécificité, permettent de déterminer un certain nombre de caractéristiques, relatives soit à la collection, soit aux préoccupations personnelles de chaque glosateur. Ainsi, entre autres, la main A se soucie assez spécifiquement de la construction syntaxique des lemmes qu'elle glose. Il en va de même pour la main C et les explications sur le *modus* du verbe, ou pour la main B et son souci sur les adverbes. Ces typologies appartiennent plutôt à la sphère de l'explication grammaticale et révèlent les objectifs pédagogiques. Ces gloses sont potentiellement le produit original de chaque glosateur, plutôt qu'une copie directe de l'*exemplar*, et ne renseignent que peu sur la *recensio* du texte.

Il n'en va pas de même pour les typologies S554, S521 et S221. Un commentaire mythologique - ici pour les lemmes *cupido*(quidam deus) et *flamen* (sacerdotum iouis significat) - peut être originaire d'une collection ou source spécifique, comme un glossaire. Pareillement pour les « quia-gloses », un commentaire explicatif introduit par *quia*(=parce que), et pour la négation d'un antonyme, qui n'apparaît pas aussi souvent que le synonyme exact d'un lemme. Encore, une comparaison entre témoins peut élucider le taux d'originalité et le « poids » qu'on attribuera à chaque glose⁵⁶.

Le poids d'une glose reflète sa (non-)trivialité :

- w=1 : une glose très triviale, qui aurait pu être et a probablement été inventée plusieurs fois indépendamment (par exemple, complète une ellipse dans une phrase) ;

^{56.} classification tirée de : E. Steinová et P. Boot, « The glosses to the first book of the Etymologiae of Isidore of Seville : a digital scholarly edition »...,Introduction

- w=2 : glose moins triviale, qui aurait pu être inventée plusieurs fois indépendamment, mais moins probable qu'une glose ayant le poids 1 ;
- w=3 : glose non triviale, qui a peu de chances d'avoir été inventée plusieurs fois, mais dont l'apparition parmi les gloses partagées indique plutôt une transmission ;
- w=4 : glose très peu triviale, qui ne peut être considérée comme ayant été inventée plusieurs fois et qui reflète donc clairement une véritable relation philologique.

Une fois que davantage de témoins ont été transcrits et qu'une collection parallèle de gloses a été établie, ces observations initiales se prêteront à une analyse plus fondée. *Quoi qu'il en soit*, cet exemple est révélateur de la manière auxiliaire dont une analyse exploratoire peut contribuer à une étude proche plus centrée, surtout lorsqu'il s'agit de données hétérogènes et à facettes multiples.

Chapitre 3

Limites et perspectives

3.1 Un modèle de SegmOnt(ation)

Certes, terminer la présentation d'un tel traitement ayant parcouru en détail tout aspects abordés est impossible. En effet, l'objectif même de cette recherche était d'ouvrir un dialogue. Par conséquent on note quelques points spécifiques embauchés qui nécessitent plus d'élaboration pour l'année prochaine.

En se penchant sur l'idée de la reconnaissance de texte plutôt que la reconnaissance des caractères¹, il reste en effet des enjeux, variables et spécifiques à un ensemble de documents, en particulier l'analyse de la mise en page et le sens de lecture de cette mise en page. Compte tenu de cette complexité des manuscrits glosés évoquée à plusieurs reprises, nous regrettons de ne pas disposer d'un modèle de segmentation qui identifierait principalement les lignes où reposent les gloses et les zones marginales où on trouve souvent des commentaires en *catena*. Ceci étant dit, il est possible, une fois que davantage de données auront été collectées, de les combiner avec les modèles entraînés pour le projet *Gallicorpora* et d'évaluer le résultats sur les témoins tels que lat.7499 (Figure. 3.1.).

^{1.} <http://www.bulac.fr/node/2491>, p.83

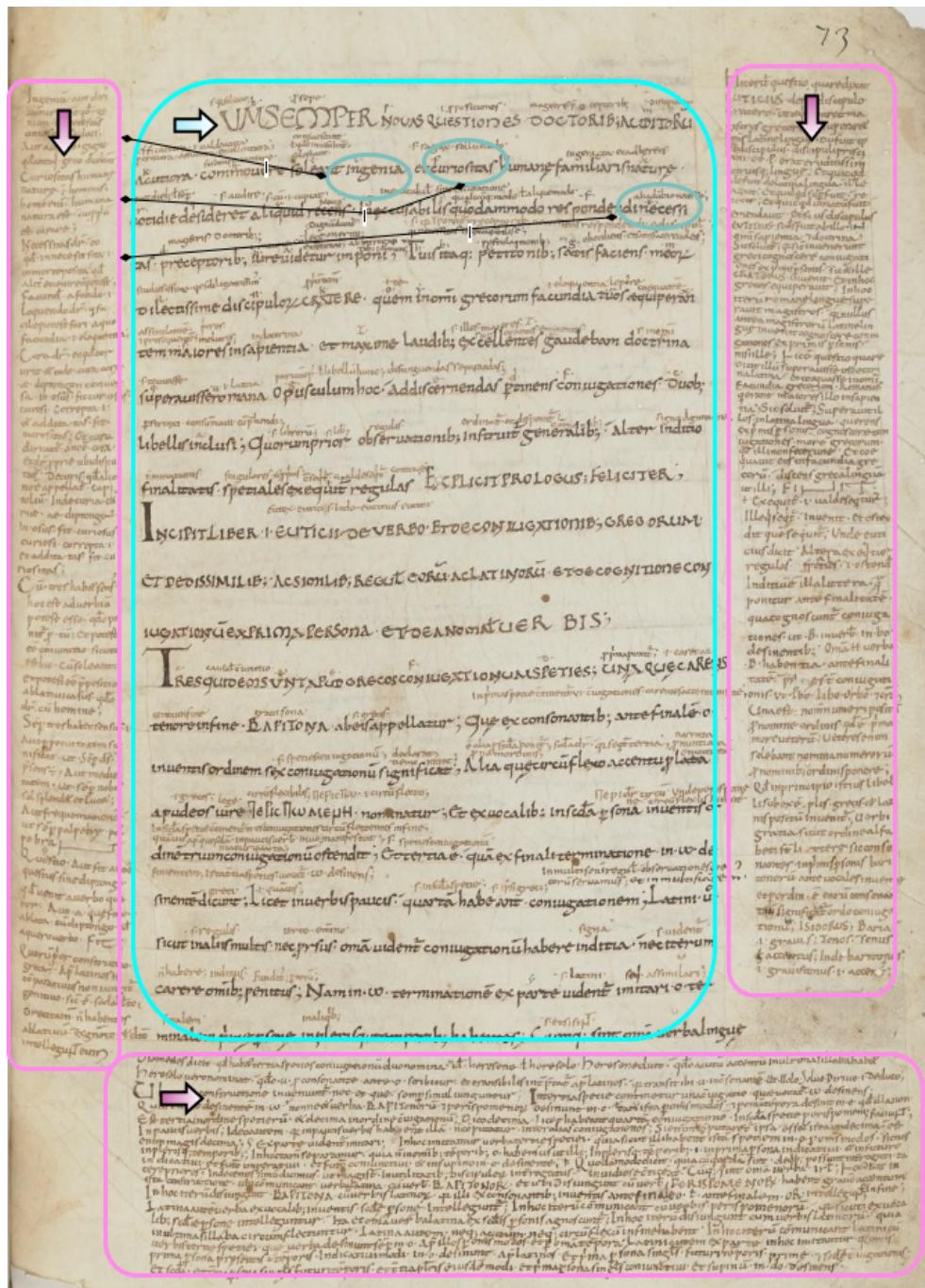


FIGURE 3.1 – La mise en page du BnF lat.7499 qui accueille en même temps le texte d'Eutychès, des gloses interlinéaires et le commentaire de Rémi d'Auxerre en marge. Les différentes couleurs marquent des Zones distinctes, les flèches le sens de lecture et les lignes noires l'interaction entre les niveaux.

En même temps, Professeur Stokes lors d'un entretien sur le modèle de segmentation, m'a proposé une solution ingénieuse pour le marquage automatique des interlignes. En utilisant les polygons, présents dans l'élément `<Polygon>` du fichier ALTO, il est possible de calculer la longueur des interlignes par rapport aux lignes principales, en soustrayant la longueur de l'interligne de celle de la ligne principale. Via un script python et lxml, nous pouvons changer le TAGREFS en l'ID d'une ligne en question dans le fichier ALTO, chaque fois qu'elle mesure une certaine longueur -

qui correspond à une InterlinearLine. Par exemple, au sein de la MainZone, si la ligne principale mesure X pixels, et la plus longue glose x pixels, toute ligne qui mesure x pixels ou moins est automatiquement marqué en tant qu'InterLinearLine. Cette méthode pourrait éventuellement être appliquée aux pages écrites uniquement en longues linges, car malheureusement la longueur des lignes à l'intérieur des colonnes coïncide forcement avec celle des gloses.

3.2 Balisage automatique du couple lemme/glose

Sans aucun doute, la partie la plus longue et la plus difficile de l'encodage a été de repérer, vérifier et baliser les couples lemmes-glosses. Etant donné les multiples attributs spécifiques et xml :id que chacun nécessite, le processus, afin d'obtenir la granularité fine souhaitée, est aussi chronophage que fastidieux. Franck Cinato² a déjà constaté qu'un encodage assez fin et à double tranchant : « Une telle approche se heurte à la limite imposée par la durée que réclame l'étape de marquage des types selon des codes informatiquement exploitables. Toutefois, la lourdeur du système se trouve compensé par l'utilité de l'intérogation. »

Parallèlement, un encodage manuel,(surtout quand on est confronté à la pression du temps et à un nombre élevé de gloses), est sujet aux *lapsus* humains. A plusieurs reprises, la vérification de l'encodage et la correction des erreurs a rendu le processus encore plus long. En effet, afin d'assurer la qualité de nos données, nous avons dû itérer trois fois la même démarche en 1) notant dans un fichier excel toutes les informations concernant les gloses, 2) encodant ces informations 3) vérifiant l'encodage. Pour toutes ces raisons, et par souci de cohérence et de validité de nos données, un repérage automatique du couple lemme-glose devrait être mis en place. Voici une solution proposée par Professeur Thibault Clérice que l'on peut appeler, vu son mécanisme, *One Transcription Does it All* (OTD). La solution consiste à annoter manuellement les lemmes et leurs gloses correspondants dès la phase de transcription, que l'on peut ensuite transformer de plusieurs manières, comme indiqué le schéma ci-dessous (Figure 3.2.) :

2. F. Cinato, *Priscien glosé...,* p.211

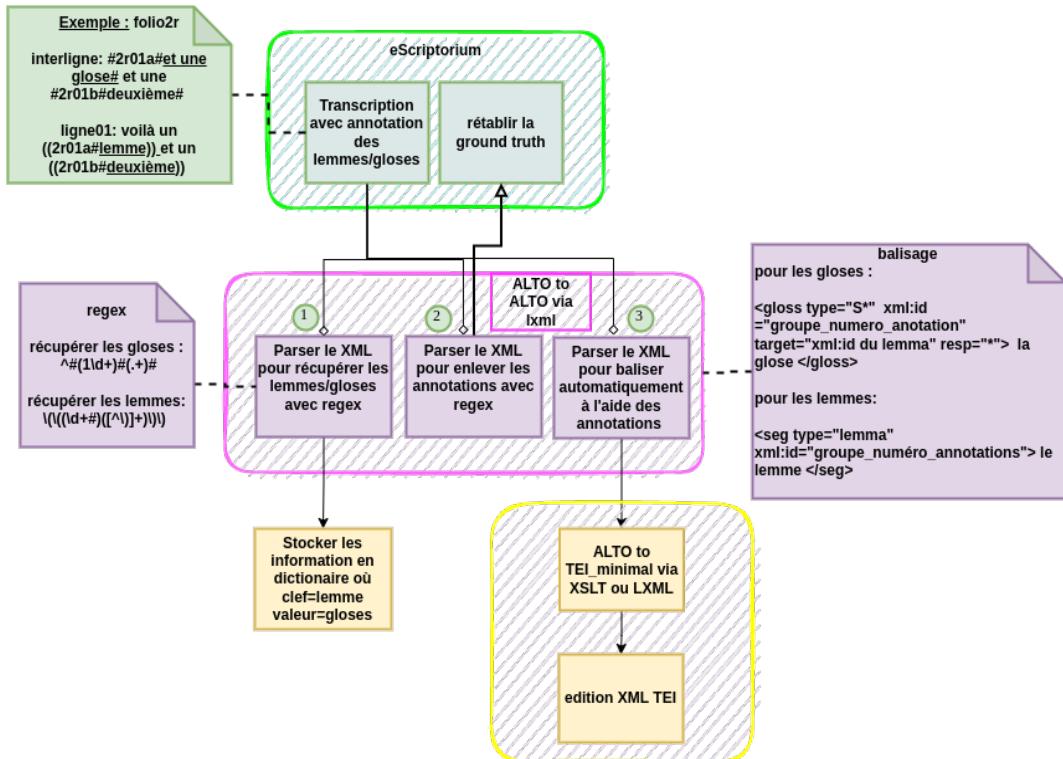


FIGURE 3.2 – Modelisation de l'approche *One Transcription Does it All*

Tout d'abord, un marquage manuel a lieu, qui consiste à entourer les lemmes et les gloses potentiels par une combinaison de signes (cf. Fig.3. note en haut à droite) qui contient les coordonnées de chaque élément (page, ligne, ordre d'apparition dans la ligne). La deuxième étape, une fois les fichiers ALTO exportés, consiste en 3 opérations distinctes (d'où la polyvalence de la transcription initiale). En premier lieu (petite boule verte (1)), l'utilisation de regex permettra de capturer les groupes d'intérêt et de les stocker en forme d'un dictionnaire où {clef = "lemme" : valeur= "glose"}. En deuxième lieu (petite boule verte (3)), la même opération peut déboucher à un balisage automatique du couple lemme-glose, en substituant le marquage manuel par la balise souhaitée (cf. figure 3.1 au milieu à droite). Ainsi, la possibilité s'ouvre de créer la base d'un encodage XML-TEI via python et lxml (partie jaune). Troisième et dernière opération (petite boule verte (2)), le rétablissement de la vérité terrain qui est perturbée, via l'enlèvement du marquage initial.

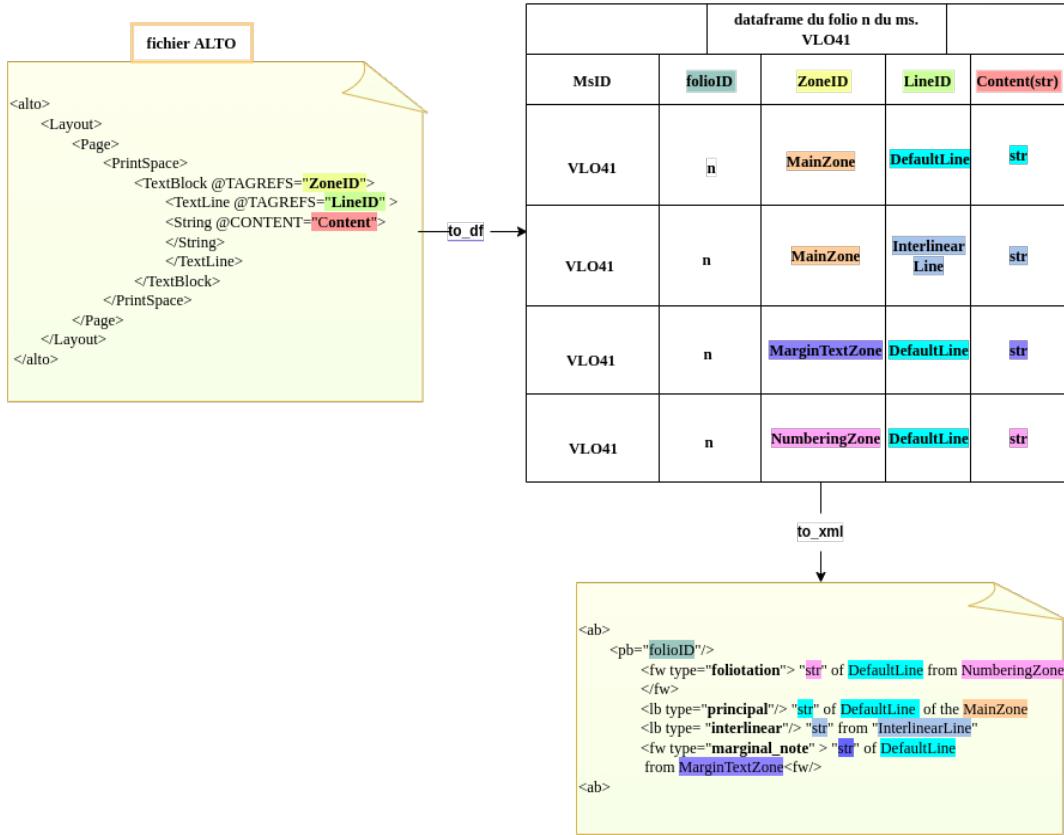


FIGURE 3.3 – Alternative : Stockage en dataframes et transformation via lxml

Une méthode alternative (Figure 3.2) qui peut complémer la précédente a été portée à notre attention par Professeur Chahan Vidal-Gorène, qui vise au stockage et à la transformation des fichiers ALTO économiques et efficaces. Du fichier ALTO exporté, porteur de l'annotation SegmOnto, on passe à un dataframe qui classe le contenu par manuscrit, folio, type de zone et type de ligne concernée, tel que présenté dans la Figure 3.3. Cette méthode s'avère pratique pour la structuration et la comparaison des témoins par l'extraction des éléments à exploiter.

Au moment où ce mémoire est en cours de rédaction, une option « Annotations » est en cours d'implémentation dans eScriptorium (pour l'instant toujours en phase beta), qui peut éventuellement donner une réponse définitive à cette question. Une fois implementée, elle permettra une annotation sémantique des morceaux de texte avec des étiquettes personnalisées, comme les lemmes et les gloses. A condition que ces étiquettes figurent dans le fichier exporté, cela devrait rendre le repérage des lemmes et gloses possible dès la phase de la transcription sans perturber la vérité de terrain.

3.3 Nouvelles perspectives - Portée du projet

Enfin, il convient d'évoquer la portée que vise un tel projet sur la tradition manuscrite d'Eutychès, à savoir la réalisation de sa propre édition de gloses. Comme il a déjà été évoqué maintes fois,

le caractère fluide, non linéaire et polygénique des gloses et la nature complexe de leur transmission (non seulement par copie d'un exemplaire à un autre, mais aussi oralement et peut-être après avoir été longtemps conservées en mémoire) ne permettent pas une édition critique au sens traditionnel. Un cadre critique alternatif impose plutôt la présentation du matériel en tant que réseau, représentant les modalités du rapport et de la similarité entre les témoins. Sur ce cadre alternatif, Evina Steinova décrit le processus³ :

In this framework, several entities (manuscript witnesses, clusters of glosses transmitted together, and chapters of the first book of the *Etymologiae*) appear as nodes. Glosses that are common to multiple of these entities appear as edges that connect them. In this manner, the users of the edition can access the corpus from several different angles and simultaneously view it in its entirety and dissected into its many constituent entities (clusters, layers of glosses in particular witnesses, and individual glosses).

Avant d'être en mesure de réaliser une édition numérique de ce type, plusieurs étapes doivent être réalisées, selon une méthodologie spécifique. Les témoins manuscrits, une fois que chaque signe a été transcrit et examiné, doivent être classés selon la typologie des gloses. En l'absence de datation précise de la copie des gloses, une typologie ne peut être réalisée qu'après avoir rassemblé des échantillons significatifs pour chacune d'entre elles. Les échantillons collectés devraient permettre d'identifier des couches de composition dans le contenu des gloses, et ainsi de définir des étapes dans le développement du corpus de gloses dans le Haut Moyen Âge. Ensuite, un manuscrit de base contenant une version étendue des gloses est choisi et les variantes du reste des manuscrits sont collationnées contre son texte. Le petit nombre de manuscrits glosés pour Eutychès permet de considérer l'ensemble du corpus. Cette activité comparative permettra la création d'un thésaurus des gloses, chaque « archéotype » de glose - une glose au sens abstrait, instanciée dans différentes versions - recevant un identifiant xml :id unique. Ce n'est qu'alors que des réseaux de gloses pourront être projetés et que des conclusions pourront être déduites sur leur contenu, en fonction de la tradition manuscrite.

3.4 Conclusions générales

L'objectif de cette étude était de démontrer, tout d'abord, au niveau théorique, l'intérêt que présente la tradition des manuscrits glosés du *De uerbo* d'Eutychès, notamment sa mise-en-page particulière et ses différents niveaux d'annotation. Dans un deuxième temps, nous avons cherché à mettre en évidence les différentes moyens dont les outils numériques s'avèrent essentiels pour la description et la manipulation adéquates des documents d'une telle nature. Plusieurs technologies ont été appliquées à notre cas d'étude, allant de la reconnaissance automatique du texte jusqu'à

3. E. Steinová et P. Boot, « The glosses to the first book of the *Etymologiae* of Isidore of Seville : a digital scholarly edition »..., Introduction, Network as a model for a Digital Scholarly Edition.

l'encodage et à la lecture distante du corpus via la visualisation de ses composants. Au même moment, les limites et les moyens d'optimiser notre approche ont été évoqués. Ce qui reste à faire et que nous tâcherons à accomplir dans l'année qui vient, suppose tout d'abord une augmentation du jeu de données, en incluant plusieurs témoins et l'optimisation du pipeline, afin de le rendre aussi cohérent et automatique que possible, tout en veillant à la qualité et à l'interopérabilité des données concernées.

Bibliographie

- CERQUIGLINI (Bernard), « Eloge de la variante : Une Histoire critique de la philologie (Paris, 1988) », *Vincent Kaufmann, L'Equivoque épistolaire (Paris, 1990)* (, 2021).
- CHAGUÉ (Alix), CLÉRICE (Thibault) et ROMARY (Laurent), « HTR-United : Mutualisons la vérité de terrain ! » (, 2021).
- CINATO (Franck), *Priscien glosé*, t. 41, 2015.
- *Les listes des grammairiens dans le haut Moyen Âge et le témoignage du Liber glossarum*, 2019.
- CINATO (Franck) et PARIS (EPHE), « Perspectives offertes par un corpus électronique de gloses sur Priscien », *Eruditio antiqua*, 3 (2011), p. 131-51.
- CLÉRICE (Thibault) et PINCHE (Ariane), *Choco-Mufin, a tool for controlling characters used in OCR and HTR projects*, version 0.0.4, sept. 2021, DOI : [10.5281/zenodo.5356154](https://doi.org/10.5281/zenodo.5356154).
- *HTRVX, HTR Validation with XSD*, version 0.0.1, sept. 2021, DOI : [10.5281/zenodo.5359963](https://doi.org/10.5281/zenodo.5359963).
- COLOMBAT (Bernard), « LES «PARTIES DU DISCOURS»(PARTES ORATIONIS) ET LA RECONSTRUCTION D'UNE SYNTAXE LATINE AU XVI^e SIÈCLE », *Langages*—92 (1988), p. 51-64.
- CONDUCHÉ (Cécile), « La mise en page d'Eutychès », dir. François Roudaut (, 2019), p. 51-68.
- DE MEYIER (Karel A), *Codices vossiani latini*, t. 16, 1973.
- DE NONNO (M), PAOLIS (P de) et HOLTZ (L), « Manuscripts and tradition of grammatical texts from antiquity to the Renaissance » (, 2000).
- DELLA CORTE (Francesco), t. 2, 1984.
- DINKOVA-BRUUN (Greti), « Text and gloss », dans *The Oxford Handbook of Latin Palaeography*, 2020.
- GABAY (Simon), CAMPS (Jean-Baptiste), PINCHE (Ariane) et JAHAN (Claire), « SegmOnto : common vocabulary and practices for analysing the layout of manuscripts (and more) », dans *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*, 2021.
- GARCEA (Alessandro), PLANQ (Clément) et CINATO (Franck), « Corpus Grammaticorum Latinorum : un projet de traitement informatique autour des grammairiens latins », *Corpus grammaticorum latinorum* (, 2010), p. 1000-1024.
- HAWK (Brandon), KARAISL (Antonia) et WHITE (Nick), « Modelling Medieval Hands : Practical OCR for Caroline Minuscule » (, 2018).
- HOLTZ (Louis), « La typologie des manuscrits grammaticaux latins », *Revue d'histoire des textes*, 7—1977 (1978), p. 247-269.

- HOLTZ (Louis), « Les manuscrits latins à gloses et à commentaires : de l'antiquité à l'époque carolingienne », dans *Il Libro e il testo*, dir. R. Raffaelli C. Questa, 1984, p. 139-167.
- « Glossaires et grammaire dans l'Antiquité », dans *Les manuscrits des lexiques et glossaires de l'Antiquité tardive à la fin du moyen âge : Actes du Colloque international (Erice, 23-30 septembre 1994)*, 1996, p. 1-21.
- JACQUART (Danielle) et BURNETT (Charles SF), *Scientia in margine : études sur les marginalia dans les manuscrits scientifiques du moyen âge à la renaissance*, t. 88, 2005.
- JANES (Juliette), PINCHE (Ariane), JAHAN (Claire) et GABAY (Simon), « Towards automatic TEI encoding via layout analysis », dans *Fantastic future 21*, 2021.
- JEUDY (Colette), *Les manuscrits de l'"Ars de uerbo" d'Eutychès et le commentaire de Rémi d'Auxerre*, 1974.
- KEIL (Henricus), *Grammatici latini ex recensione Henrici Keilii.-Lipsiae, BG Teubner 1857-1880*, t. 1-5, 1857.
- KIESSLING (Benjamin), TISSOT (Robin), STOKES (Peter) et EZRA (Daniel Stökl Ben), « eScriptorium : An open source platform for historical document analysis », dans *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, IEEE, 2019, t. 2, p. 19-19.
- KUHRY (Emmanuelle), « Medieval Glosses as a Test Subject for the Building of Tools for Digital Critical Editions », *Journal of the Text Encoding Initiative*-13 (2020).
- « Vers une édition électronique de la glose d'Oxford », *Humanités numériques*-2 (2020).
- LAW (Vivien), *The insular Latin grammarians*, 1982.
- « From Aural to Visual : Medieval representations of the word », *Grammar and Grammarians in the early Middle Ages* (, 1997), p. 250-259.
- *Grammar and grammarians in the Early Middle Ages*, 1997.
- LAW (Vivien A), « Late Latin grammars in the Early Middle Ages : a typological history », *Historiographia linguistica*, 13-2-3 (1986), p. 365-380.
- LINDEMANN (Friedrich), *Corpus grammaticorum latinorum veterum*, t. 3, 1833.
- LOMANTO (Valeria), « Eutiche », dans *Enciclopedia virgiliana*, 1985.
- « A concordance to Keil's Latin grammarians », *Computers and the Humanities*, 24-5 (1990), p. 427-435.
- LOMANTO (Valeria) et MARINONE (Nino), *Index grammaticus : an index to the Latin grammar texts*, t. 81, 1990.
- MARTORELLI (Luca), « Le citazioni in Eutiche », *Revue de philologie, de littérature et d'histoire anciennes*, 91-2 (2017), p. 55-88.
- MONELLA (Paolo), « Towards a digital model to edit the different paratextuality levels within a textual tradition », *Digital Medievalist*, 4 (2008).
- « A digital critical edition model for Priscian », *M. Pade (ed. by), Philology Then and Now : History, Role, and New Directions [im Erscheinen]* (, 2019).

- PIERAZZO (Elena), « A rationale of digital documentary editions », *Literary and linguistic computing*, 26-4 (2011), p. 463-477.
- *Digital scholarly editing : Theories, models and methods*, 2016.
- PIERAZZO (Elena) et STOKES (Peter A), « Putting the text back into context : a codicological approach to manuscript transcription », dans *Kodikologie und Paläographie im digitalen Zeitalter 2—Codicology and Palaeography in the Digital Age 2*, 2011, p. 397-430.
- ROBINSON (Peter) et VAN VLIET (HTM), « What is a critical digital edition », *Variants*, 1 (2001), p. 43-62.
- SABBADINI (Remigio), *Opere minori : Classici e umanisti da codici latini inesplorati*, t. 87, 1995.
- SCHAD (Samantha), *A lexicon of Latin grammatical terminology*, 2007.
- STEINOVÁ (Evina), *Notam superponere studui : The use of technical signs in the early Middle Ages*, thèse de doct., Utrecht University, 2016.
- STEINOVÁ (Evina) et BOOT (Peter), « The glosses to the first book of the *Etymologiae* of Isidore of Seville : a digital scholarly edition », (2021).
- VLACHOU-EFSTATHIOU (Malamatenia), *Voss.Lat.O.41 - Eutyches "de uerbo" glossed*.
- ZETZEL (James EG), *Critics, compilers, and commentators : An introduction to Roman philology, 200 BCE-800 CE*, 2018.

Table des figures

2.1	Modelisation du pipeline suivi. Trois niveaux de traitement successifs, de la transcription à la structuration et la visualisation des données.	12
2.2	(a) et (b) Disposition des zones SegmOnto afin de rétablir le sens de lecture. (c) DefaultLines en rose, InterlinearLines en orange.	16
2.3	Transformation des fichiers ALTO à un fichier XML-TEI minimal avec intégration de l'ontologie SegmOnto	20
2.4	Rappel du pipeline : Modelisation des opérations de transformation des données brutes en données structurées jusqu'à l'encodage complet (axe horizontale)	30
2.5	Dans le premier graphique, une ligne correspond à une main en traçant sa contribution per folio. Le deuxième prend les mêmes variables et en donne la densité de la contribution de chaque main.	32
2.6	Somme des gloses écrites par chaque main. "Mg" vaut pour Main de glosateur. . .	32
2.7	Negatif du folio 6v avec échantillon des <i>notae</i> de la main C	33
2.8	Circular barplot pour la distribution hiérarchique du type de gloses utilisé par main.	35
3.1	La mise en page du BnF lat.7499 qui accueille en même temps le texte d'Eutychès, des gloses interlinéaires et le commentaire de Rémi d'Auxerre en marge. Les différentes couleurs marquent des Zones distinctes, les flèches le sens de lecture et les lignes noires l'interaction entre les niveaux.	40
3.2	Modelisation de l'approche <i>One Transcription Does it All</i>	42
3.3	Alternative : Stockage en dataframes et transformation via lxml	43