Données et modèles pour le traitement des documents en néolatin : le cas Lambert Daneau

Floriane Goy[1], Noemi Schürmann[2], Benjamin Manig [2], Matteo Colombo[1], Ueli Zahnd[1], and Stefan Krauter[2]

[1] Institut d'histoire de la Réformation, Université de Genève, Suisse
[2] Université de Zürich, Suisse

Abstract

This article presents the construction of a corpus of sixteenth-century commentaries on the Epistles of Paul, based on the digitization of numerous printed works in Neo-Latin. As this subtype of Latin is still underrepresented in existing datasets, it required the development of specific resources for training suitable models. The prepared data and models for ATR post-correction and lemmatization are described here to enable systematic digital exploitation of the historical material.

Mots-clés: philologie latine, néolatin, analyse de mise en page, reconnaissance automatique de texte, normalisation linguistique, lemmatisation
Keywords: latin philology, neo-latin, layout analysis, automatic text recognition, linguistic normalisation, lemmatisation

1 Introduction

Avec les possibilités ouvertes par la numérisation massive des manuscrits et des imprimés conservés dans les bibliothèques et les institutions patrimoniales, la capacité à traiter de grands ensembles de données s'impose comme une urgence pour nombre de projets. L'ambition générale du projet FNS sur l'exégèse de Paul au XVI^e^ siècle [38] est ainsi de numériser et analyser les commentaires réformés des Épîtres de Paul en utilisant les imprimés d'époque, rarement édités, disponibles en ligne, et notamment ceux conservés en Suisse [1] et en Allemagne [2]. Une base de données, *Reformation Readings of Paul* [3] propose également une première liste des ouvrages avec un lien vers leurs exemplaires digitalisés lorsqu'ils existent.

Les commentaires exégétiques latins du XVI^e^ siècle constituent une masse documentaire considérable, encore trop peu étudiée malgré son importance. Une comparaison à grande échelle de ces textes est ainsi susceptible de fournir un éclairage nouveau sur la complexité des dynamiques intellectuelles à l'œuvre durant cette période charnière entre le Moyen Âge et l'Époque moderne. Pour mener à bien une telle entreprise, la lecture distante [28] offre une approche particulièrement pertinente, notamment à travers des méthodes d'analyse linguistique telles que la modélisation de sujets (*topic modelling*) [33]. La mise en œuvre de ce type d'analyse requiert toutefois, dans un premier temps, l'acquisition des textes conservés dans les imprimés anciens par reconnaissance automatique de texte (*Automatic Text Recognition*, ATR), puis leur enrichissement au moyen d'outils de Traitement Automatique des Langues (TAL), en particulier la lemmatisation. Cette communication

Floriane Goy, Noemi Schürmann, Benjamin Manig , Matteo Colombo, Ueli Zahnd, and Stefan Krauter. "Données et modèles pour le traitement des documents en néolatin: le cas Lambert Daneau." *Actes de la Conférence Humanistica*, éd. par Serena Crespi, Simon Gabay, Martin Grandjean, Ariane Pinche, Marie Puren et Léa Saint-Raymond. Vol. 4. Anthology of Computers et the Humanities. 2026, 120–133. https://doi.org/10.63744/5TcizCXUUTmJ.

1. `https://www.e-rara.ch`.
2. `https://www.digitale-sammlungen.de`.
3. `https://ihr-num.unige.ch/rrp`.

Auteur	Cote	Date	DLA	ATR	TAL
J. Lefèvre d'Étaples	Regensburg, Staatliche Bibliothek, 999/2 Script.801	1512	3	-	-
J. Lefèvre d'Étaples	Regensburg, Staatliche Bibliothek, 999/2 Script.801	1512	41	-	-
J. Bugenhagen	München, Bayerische Staatsbibliothek, Res/Exeg. 309 b Beibd.3	1524	18	-	-
M. Bucer	München, Bayerische Staatsbibliothek, Polem. 408 Beibd.2	1527	19	395	-
T. Cajetan	München, Bayerische Staatsbibliothek, 2 Exeg. 610	1531	6	-	-
T. Cajetan	Augsburg, Staats- und Stadtbibliothek, 2 Th Ex 424	1532	-	58	-
Anonyme	Basel, Universitätsbibliothek, UBH FG VIII2 16 :7	1533	24	-	-
K. Megander	München, Bayerische Staatsbibliothek, Exeg. 700 m	1534	-	273	-
H. Bullinger	Zürich, Zentralbibliothek, VD 16 B 5144	1536	32	-	-
M. Bucer	Regensburg, Staatliche Bibliothek, 999/2 Script.662	1536	19	-	-
C. Pellicanus	Zürich, Zentralbibliothek, III B 14 | G	1539	11	-	-
J. Calvin	Bibliothèque de Genève, Bb 1493 (2)	1548	15	-	-
L. Daneau*	Bibliothèque de Genève, Cti 1753 BGE S 22877	1577	-	4391	18 329
B. Aretius	München, Bayerische Staatsbibliothek, Exeg. 53 Beibd.1	1580	163	354	-
A. Hyperius	Zürich, Zentralbibliothek, C 85 | G	1582	12	-	-
Total			363	5471	18 329

TABLEAU 1 – Corpus de travail. DLA : Document Layout Analysis (§ 4.1), total en nombre de pages ; ATR : Automatic Text Recognition (§ 4.2) : total en nombre de lignes ; TAL : Traitement Automatique des Langues (§ 4.4), concerne ici l'annotation linguistique : total nombre de tokens.

présente les premiers résultats obtenus à l'issue de ces deux étapes, en amont d'une exploitation théologique des données qui sera développée ultérieurement.

La difficulté liée à la constitution du corpus paulinien tient à son caractère inédit : si le latin figure parmi les langues les mieux pourvues en ressources informatiques, le néolatin du XVI^e^ siècle demeure largement sous-doté. Or cet état de langue présente des spécificités importantes, tant sur le plan graphique que lexical. Il impose dès lors la mise au point de ressources *ad hoc*, qui doivent en outre rester interopérables avec celles élaborées pour le latin classique. C'est à la résolution de ce double enjeu que s'attache notre travail.

Le travail devant nécessairement trouver un point de départ, c'est un commentaire de l'épître à Timothée rédigé par Lambert Daneau (1530-1595) [11] qui a concentré l'essentiel de nos efforts. Rédigé en latin au XVI^e^ siècle par un théologien calviniste, ce texte nous a en effet semblé particulièrement représentatif des imprimés de notre corpus, tant du point de vue linguistique que formel (usage des abréviations, casse typographique, etc.). Relativement peu connu, ce commentaire offre en outre l'occasion de remettre en lumière l'importance théologique de son auteur [37].

2 État de l'art

De nombreux projets se sont intéressés aux corpus de textes en langue latine de l'Antiquité, qu'il s'agisse d'initiatives lancées par le secteur privé comme la *Library of Latin Texts* de Brepols[4], dont le corpus couvre également les textes en médiolatin, ou de projets publics comme le portail de textes gréco-latins *Perseus*[5]. En parallèle de ces grands projets généralistes, on trouve des projets traitant de thématiques plus spécialisées comme *Musisque Deoque*[6], un répertoire dédié à la poésie métrique. Pour les documents en latin médiéval, là encore il existe des outils privés, comme une version numérique de la *Patrologia Latina* de Migne[7], ou des initiatives plus ambitieuses et non commerciales comme le *Corpus Corporum*[8]. Le néolatin, lui, n'a en revanche pas eu la chance de connaître des initiatives de cette ampleur. Il fait cependant l'objet de projets plus modestes regroupant des corpus régionaux, à l'instar de *Croala*[9] qui rassemble des textes originaires des Balkans ou la *Database of Nordic Neo-latin Literature*[10] au Danemark.

Avec le développement des dernières technologies, l'ATR est devenu un passage obligé pour la constitution rapide et efficace de grands corpus. Des premiers modèles de grande qualité ont été publiés pour les manuscrits médiévaux [10 ; 30] ou les imprimés en caractères gothiques [34], mais ceux pour les imprimés modernes en caractères romains [21 ; 29] ne répondent encore qu'imparfaitement aux besoins des spécialistes de néolatin. N'ayant presque jamais vu de données dans cette dernière langue, les modèles disponibles commettent encore un grand nombre d'erreurs de transcription (diacritiques, etc.).

Une fois les données extraites avec l'ATR, le chercheur se trouve confronté au problème de la normalisation. L'ATR fournit en effet une transcription graphématique du texte. Or, la lemmatisation requiert un vêtement graphique qui doit être minimalement lissé (abréviations, soudures, etc.). Il faut donc normaliser le texte, c'est-à-dire l'aligner sur une norme qui facilite le travail de la machine. Si des stratégies de normalisation de la langue pour le vernaculaire ont été développées [1], il n'y a rien de tel pour le latin du XVI^e^ siècle.

L'annotation linguistique (lemmes, parties du discours, morphologie) pose aussi problème. Si le latin classique [5], l'ancien français [3] ou le français moderne [18] ont déjà des jeux de données d'entraînement et des modèles facilement accessibles [6], on note l'absence de solution pour le latin du XVI^e^ siècle. Ce dernier présente cependant des particularités lexicales et graphiques évidentes, si l'on pense à l'omniprésence et l'importance du lexique chrétien, notamment pour un corpus de commentaires théologiques.

3 Données

Outre les travaux de Lambert Daneau, les données mobilisées pour les différentes expériences présentées ici proviennent d'imprimés publiés entre 1512 et 1582, sélectionnés sur recommandation des doctorants·e·s du projet. Ces imprimés contiennent tous des commentaires aux *Épîtres de Paul* rédigés par des réformateurs et présentent une grande variation de longueur, allant d'environ 200 pages à près de 500 pages. Les facsimilés numériques ont été obtenus via la bibliothèque numérique E-Rara[11] ou le Münchener Digitalisierungszentrum[12]. Ce corpus (tab. 1) est naturellement destiné à évoluer au fil du projet.

4. https://www.brepols.net/series/LLT-O.
5. http://www.perseus.tufts.edu/hopper.
6. https://www.mqdq.it/public.
7. https://www.proquest.com/patrologialatina.
8. https://mlat.uzh.ch.
9. https://www.ffzg.unizg.hr/klafil/croala/
10. https://cdnl.dk/dbnnl/dbnnl_search.htm
11. https://www.e-rara.ch.
12. https://www.digitale-sammlungen.de.

4 Chaîne de traitement

Notre chaîne de traitement comporte ainsi trois étapes principales (fig. 1) : l'extraction des données — incluant l'analyse de la mise en page (§ 4.1) et l'ATR (§ 4.2) —, suivie de leur normalisation et de leur annotation linguistique (§ 4.4). Les outils existants pour la préparation des données, qu'il s'agisse d'eScriptorium [27] ou de Pyrrha [8], sont heureusement compatibles avec le néolatin. Les données au format XML-ALTO produites par eScriptorium sont converties en XML-TEI. [13]

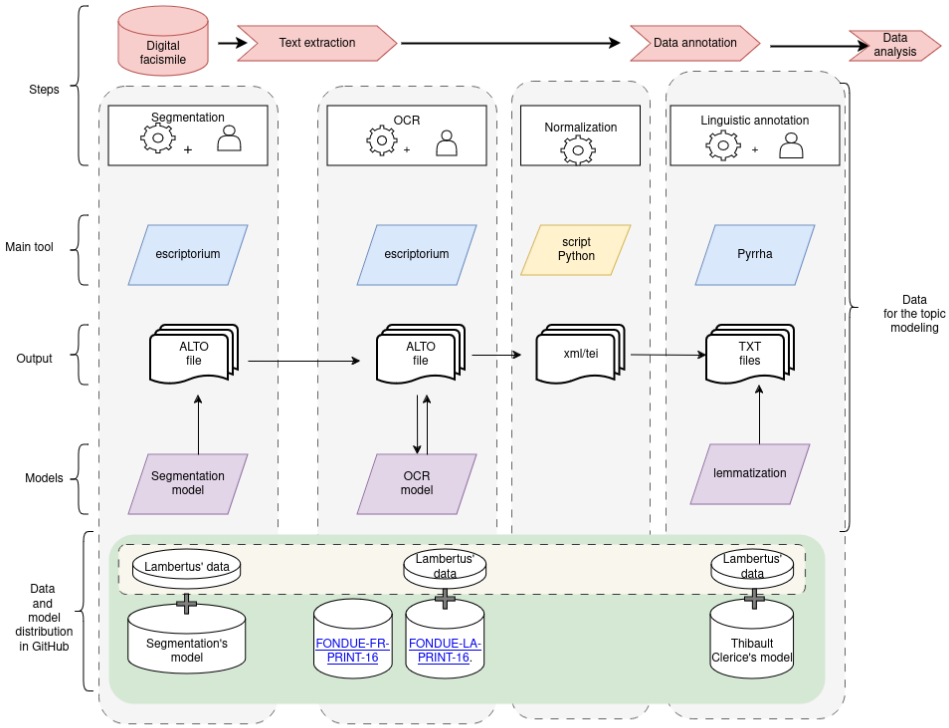

FIGURE 1 – Schéma décrivant les 4 étapes principales du traitement des données : la segmentation (§ 4.1), l'ATR (§ 4.2), la normalisation (§ 4.3), l'annotation linguistique (§ 4.4).

4.1 Analyse de mise en page

L'analyse de la page a été l'un des enjeux majeurs du projet. Pour ce faire, nous avons d'abord bénéficié d'un modèle de segmentation pour les imprimés français du XVI^e^ siècle [24 ; 34], puis du modèle de segmentation large développé dans le projet LaDaS [7].Pour finir, nous avons adapté le modèle LaDaS aux spécificités propres à la mise en page des commentaires exégétiques latins du XVI^e^ siècle en fine-tunant un nouveau modèle `Layout-16th-Print-Lat`[23] à partir de ce dernier [14].

4.1.1 Données d'entraînement

Nos données (tab. 1) sont distribuées dans un dépôt GitHub préparé selon les recommandations du projet *HTR-United* [4] : *HTR-Corpus-A* [15]. Pour garantir l'interopérabilité des données, nous avons

13. tous les scripts utilisé pour le traitement des données sont disponibles sur le dépôt suivant `https://github.com/16thExegesisDH/PipeLineThm`.
14. `https://github.com/ponteineptique/yaltai`
15. `https://github.com/16thExegesisDH/HTR-Corpus-A`

recours au vocabulaire de SegmOnto [20] tel que revu par le projet LADaS [7] (fig. 2)[16]. Nos documents ne posent pas de problèmes particuliers en matière de mise en page. La préparation des données est faite avec eScriptorium [27], dont l'Université de Genève héberge une instance [16].

4.1.2 Modélisation de la page

Notre objectif principal est une restructuration du document (texte vs paratexte, différentes parties et sous-parties) ainsi que la reconnaissance aussi automatique que possible des zones de titre intermédiaire (`MainZone-Head`) contenant très fréquemment les versets bibliques. Dans ce but, neuf types de zones ont été retenues (§ A).

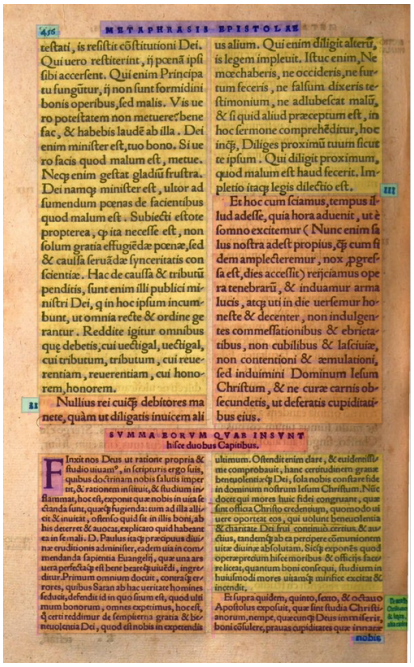

FIGURE 2 –
Exemple de données d'entraînement.

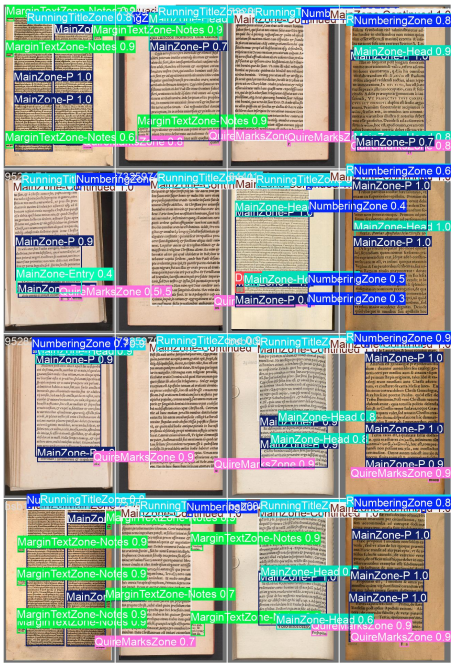

FIGURE 3 –
Prédiction effectuées par notre modèle.

4.1.3 Entraînement et évaluation du modèle

Le modèle de segmentation que nous avons entraîné a permis d'automatiser la tâche d'annotation des images et ainsi d'accélérer le travail. En dépit de quelques faiblesses persistantes, il présente les performances suivantes :

— La matrice de confusion (fig. 4) met en évidence une très bonne reconnaissance des éléments récurrents de la mise en page (`RunningTitle`, `MainZone-Continued`, `MainZone-P`, `QuireMarksZone`, `DropCapitalZone`, `MarginText-Notes`).

— `MainZone-Head` est fréquemment correctement identifiée avec un score de 0,88, marquant une amélioration notable par rapport au modèle LaDaS, qui, dans le cas de nos imprimés, la confondait avec d'autres zones ou ne la détectait pas.

— La principale faiblesse du modèle concerne la `NumberingZone`. En raison de son ambiguité, la zone regroupe indistinctement la numérotation de page et la numérotation marginale du commentaire biblique (fig. 2) ce qui engendre une confusion.

— Les éléments non reconnus relèvent majoritairement de données exclues du modèle (page de titre, index) ou insuffisamment représentées dans les données d'entraînement (par ex. `GraphicZone`).

16. Le vocabulaire LADaS ayant évolué entre temps, notamment pour `-P-Continued` devenu `Continued`, le modèle a été produit en utilisant la première version, mais les données ont été corrigées dans le sens de la nouvelle version.

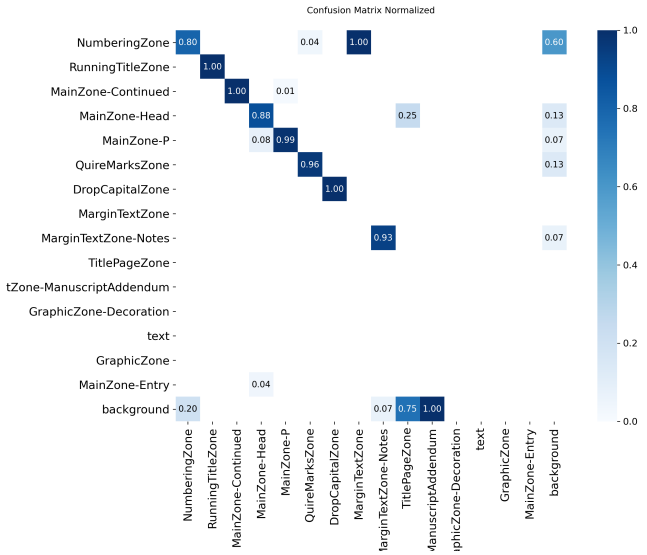

FIGURE 4 – Précision de la segmentation en fonction des zones.

4.2 Reconnaissance automatique de caractères

Au début de notre projet, nous avions à notre disposition le modèle du projet *Gallicorpora* [31], intégralement entraîné sur des données en français. Malheureusement, les premiers tests ont démontré d'importantes lacunes sur deux points : la transcription des abréviations et l'utilisation d'accents (fig. 5).

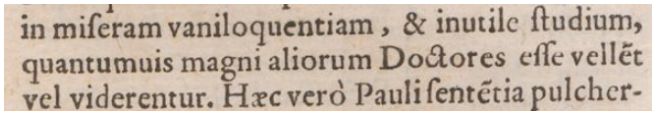
in miſeram vaniloquentiam, & inutile ſtudium, quantumuis magni aliorum Doctores eſſe vellẽt vel viderentur. Hæc verò Pauli ſentẽtia pulcher-

FIGURE 5 – Exemple d'abréviation latine (‹vellẽt› et ‹sentẽtia› avec un tilde sur le *e* pour *en*) et d'accent typique du néolatin (‹verò› doté d'un accent grave sur le *o*.) LAMB.DAN. *Ep. 1 Tim.* I,16.

4.2.1 Données d'entraînement

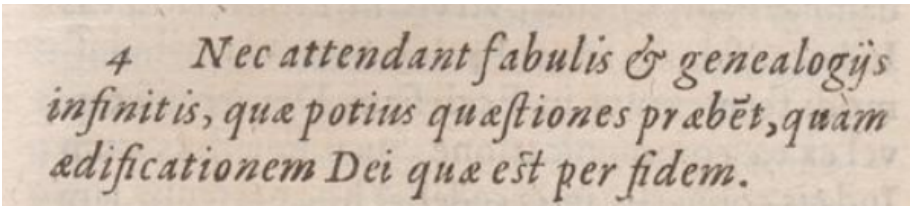
4 Nec attendant fabulis & genealogÿs infinitis, quæ potius quæſtiones præbẽt, quàm ædificationem Dei quæ eſt per fidem.

FIGURE 6 – L'italique est une typographie régulièrement employée pour distinguer visuellement le verset biblique de son commentaire. Cf. par ex. LAMB.DAN. *Ep. 1 Tim.* I,9.

Si l'essentiel de nos données provient du commentaire de Lambert Daneau, nous avons recours à des échantillons provenant de pages prises de manière aléatoire dans les différents imprimés de notre corpus de travail (tab. 1).

Nos données comprennent, outre les documents que nous avons préparés, d'autres imprimés néolatins transcrits par Marie Jeannot-Tirole, provenant de son travail de thèse sur Ioannes Sapidus [17]. Les œuvres de ce dernier, des poèmes plutôt écrites en italique [35], permettent d'augmenter

17. `https://mariejeannot.gitpages.huma-num.fr/sapidus`.

la quantité de ce type de caractère, dont l'usage restait rare (fig. 6) dans le corpus d'ATR que nous avions choisi.

Nos données ainsi que celles de M. Jeannot-Tirole, avec qui nous partageons les mêmes règles de transcription, sont distribuées dans un dépôt GitHub préparé selon les recommandations du projet *HTR-United* [4] : `FONDUE-LA-PRINT-16` [19].

Afin d'augmenter la quantité de données, nous avons aussi eu recours à des données provenant d'imprimés en langue française [14], obéissant également aux mêmes règles de transcription que les nôtres. Ce *dataset* représente environ le tiers du total des données pour le XVI^e^ s. (fig. 7).

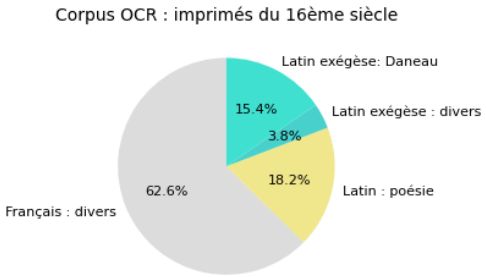

FIGURE 7 – Origine de la vérité de terrain.

4.2.2 Règles de transcription

Les règles de transcription présentées dans le projet CREMMA [9] couvrent l'essentiel des cas complexes. On y trouve en effet des recommandations pour les abréviations du latin médiéval, dont le système néolatin est une simplification. En dehors du phénomène d'accentuation, le néolatin ne présente donc pas de graphie nouvelle par rapport à celles des manuscrits médiévaux. Le seul caractère ajouté à notre transcription est ainsi le pied de mouche que les imprimeurs emploient pour indiquer les débuts des cahiers. On gardera à l'esprit qu'un corpus plus large donnera certainement à voir d'autres de ces signes typographiques.

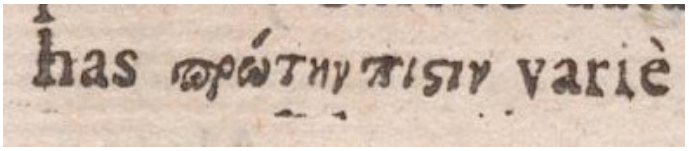

FIGURE 8 – Exemple d'un passage en grec transcrit sans ligature, transcrit *has πρώτην πιστιν variè*.

Plus épineuse est la question des langues étrangères, comme le grec et l'hébreu, qui font en effet leur apparition dans les commentaires bibliques du XVI^e^ siècle. Dans le cas des commentaires pauliniens, l'hébreu est employé de manière anecdotique, il n'a donc pas été considéré dans cette étude. En revanche, les exégètes citent fréquemment le vocabulaire du texte source des épîtres : le grec. La transcription de cette seconde langue revêt donc un intérêt particulier pour la compréhension du propos du commentaire. Or, le grec des imprimés se décline à travers un système de ligature complexe [36] qui nécessiterait à lui seul l'entraînement d'un modèle. Pour des raisons d'efficacité, il a été décidé de ne pas tenir compte de ces ligatures.

4.2.3 Entraînement et évaluation du modèle

Nous entraînons et évaluons trois modèles différents avec Kraken [26].

— `Gallicorpora+`[18] le modèle entraîné sur des imprimés anciens pour le projet `Gallicorpora`, qui n'a pas vu de vérité de terrain en (néo)latin ;
— `gallicorpora_ajuste`[22] : version ajustée de `Gallicorpora+` sur les données du XVI^e^ s. précédemment décrites ;

18. Modèle de Gallicorpora+.

— `catmusPrint_ajuste` : version ajustée du modèle CATMuS *print* [21] sur les données du XVI^e s. précédemment décrites et des données du XVII^e s. [13] [19].

Pour évaluer ces modèles, nous utilisons un double corpus hors domaine décrit dans le tab. 3. Nous utilisons deux jeux de données différents : le premier contient des extraits de documents très proches de notre corpus de travail (test 1 : commentaires pauliniens), et le second un document qui est toujours un commentaire biblique, mais dont le thème et la facture sont différents (test 2 : commentaires bibliques).

Corpus	**Métrique**	`Gallicopora`	`gallicorpora_ajuste`	`catmusPrint_ajuste`
Test 1	Acc	97.95	98.96	**99.14**
	WAcc	88.32	93.06	**94.51**
Test 1+2	Acc	97.39	98.56	**98.95**
	WAcc	86.19	91.11	**93.35**

TABLEAU 2 – Résultats des 3 modèles sur les deux jeux de test. Les meilleurs résultats sont indiqués en gras. Le calcul de la WAcc est fait avec kamiCLI.

Sans grande surprise, si les résultats sont satisfaisants pour les trois modèles, le modèle `catmusPrint_ajuste` obtient les meilleurs scores, notamment avec l'ajout du jeu de test 2. Si le gain au niveau de la *character accuracy* (CAcc) est relativement modeste, celui au niveau de la *word accuracy* (WAcc) est plus net, justifiant l'entraînement de nouveaux modèles en dépit de scores *a priori* corrects au premier abord.

Corpus	**Cote**	**Pages**	**Lines**
Test 1	BGE Cti 1753 BGE S 22877	2	72
	BCUL, AZ 8515 (4)	2	61
	ZB, C 283,2	2	51
Test 2	BGE Cta 3110 (1)		
	BGE Bb 1049 (1)	2	88
Total		**8**	**272**

TABLEAU 3 – Corpus de données *out-of-domain* pour le test d'ATR.

4.3 Normalisation linguistique

De récentes études ont démontré la supériorité des approches neuronales pour la normalisation linguistique [1 ; 15]. Les imprimés en néolatin présentent cependant une relative stabilité tant dans leurs pratiques graphiques que dans le système abréviatif : nous avons donc choisi une approche à base de règles, en utilisant des expressions régulières [20].

4.3.1 Méthode

La normalisation est ici comprise comme un alignement de la transcription sur les normes régissant le latin classique, la plupart des outils disponibles étant prévus pour traiter ce type de texte. Elle concerne en particulier :

— le système abréviatif : résolution des tildes… ;
— les accents et les cédilles : suppression des accents typiques des imprimés en néolatins… ;
— les ligatures : passage de ‹æ›/‹e› cédillé à ‹ae›, de ‹œ› à ‹oe›…
— les allographes : passage du *s* long (‹ſ›) au *s* rond ;

19. Modèle entraîné par S. Gabay, Th. Clérice, et al. [17].

20. Le script est disponible en ligne :`https://github.com/16thExegesisDH/PipeLineThm/tree/main/PYTHON/normalisation` .

Notre script permet un lissage de la variation graphique par rapport au latin classique, mais pas de corriger des évolutions plus lourdes, du type *auctor* (latin classique) → *author* (néolatin) ou *saeculum* (latin classique) → *seclum* ou *seculum* (néolatin).

4.3.2 Résultats

La normalisation est stockée dans une balise `<reg>` à côté du texte extrait par le moteur ATR, conservé dans un `<orig>`.

```
<lb corresp="#f2013098_line_26_ligne_8"/>
    choice>
        <orig>affectum animi. Sanæ, doctrinæ opponitur tum</orig>
        <reg type="normalized">affectum animi. Sanae, doctrinae opponitur tum</reg>
    </choice>
<lb corresp="#f2013098_line_27_ligne_9"/>
    <choice>
        <orig>quod ſuprà dixit eam doctrinã Diabolicam eſſe:</orig>
        <reg type="normalized">quod supra dixit eam doctrinam Diabolicam esse:</reg>
    </choice>
```

4.4 Annotation linguistique

La préparation des données pour entraîner les modèles d'annotation linguistique a été effectuée avec *Pyrrha* [8].

4.4.1 Données d'entraînement

Nous avons contrôlé 18 329 tokens [21], tous tirés des deux premiers chapitres du commentaire de Lambert Danneau (§1), soit environ septante pages situées au début de l'œuvre (tab. 4).

Parties	Pages	Lemmes
Préface	24	4 781
Chapitre I	37	9 477
Chapitre II	15	4 071
total	76	18 329

Tableau 4 – Corpus pour l'annotation linguistique.

4.4.2 Référentiel, cas particuliers

Les lemmes proviennent du dictionnaire de Forcellini [12], afin de de conserver une interopérabilité avec le travail de Th. Clérice [5]. L'utilisation de *Pyrrha* permet de garantir une qualité minimale de l'annotation (contrôle des valeurs possibles en POS, liste de lemmes possibles…).

La qualité de la prédiction proposée par le modèle LASLA distribué par *Pie extended* [6] est globalement de grande qualité. L'essentiel des problèmes concerne :

— le vocabulaire propre au latin chrétien et néolatin : *papista*, *excommunicatio*… ;
— les noms bibliques : *Ezechias*… ;
— les noms des penseurs chrétiens : *Augustinus*… ;
— les variantes néolatines de mots classiques : *author*, *seculus*… ;
— Les abréviations des différents livres bibliques : *Apocalyp* pour *Apocalypsis*…

Les abréviations de livres bibliques sont le problème majeur de notre corpus exégétique, car elles ne sont pas standardisées. Par exemple, le nom de l'Évangile de Matthieu est abrégé de trois manières différentes : ‹Math›, ‹Matth›, ‹Matt›. Nous avons choisi de regrouper ces trois tokens derrière le lemme *Matthaeus* tel que l'orthographie le dictionnaire de Forcellini. À chaque fois que nous rencontrons une abréviation biblique, nous en gardons la trace dans un fichier csv, ce qui nous permet de tenir à jour leur liste à compléter avec l'avancée du projet [22].

21. https://github.com/FourbeFlo/Lemmatization/blob/main/Lambertus_corrected.tsv.
22. https://github.com/FourbeFlo/Lemmatization/blob/main/dictionaries%20and%20abbreviations/livre-biblique.csv.

4.4.3 Résultat

Les données d'entraînement pour un méga modèle n'étant pas publiques, nous nous sommes tournés vers CLTK (Classical Latin Toolkits)[25], une bibliothèque Python de NLP pour les langues pré-modernes qui intègre notamment les modèles neuronaux de Stanza [32][2] pour l'annotation linguistique. Bien que n'ayant pas la précision que l'on pourrait espérer avec un annotateur linguistique entraîné spécifiquement pour notre corpus, la grande variété de textes latins intégrés par CLTK permet des résultats satisfaisants lors de la lemmatisation des textes comme le démontre les premiers essais de *topic modeling* (fig. 9).

FIGURE 9 – Résultat de topic modeling effectué sur LAMB.DAN. *Ep. 1 Tim.* ; le *Topic* 7 concernant la notion de femme : *mulier*.

5 Conclusion

Le travail présenté constitue une description de notre chaîne de traitement, laquelle pourrait sans doute être encore améliorée, notamment par l'ajout de nouvelles données *gold* pour renforcer la performance des différents modèles. Le corpus définitif n'étant pas encore entièrement stabilisé, il est possible que nous rencontrions à l'avenir des cas présentant des variations de mise en page, de typographie ou d'autres caractéristiques absentes des données actuellement disponibles. Les données que nous avons préparées ont néanmoins déjà permis de consolider certains modèles, tels que ceux dédiés à l'analyse de mise en page ou à l'ATR, et continueront de jouer un rôle essentiel dans des tâches comme l'annotation linguistique. En ce sens, elles ont contribué à faciliter les études computationnelles sur le néolatin, une langue encore insuffisamment explorée sur le plan numérique.

Financement

Les recherches ont été menées dans le cadre du projet *Paulusauslegung des 16. Jahrhunderts* (FNS n° 207696).

Références

[1] Bawden, Rachel, Poinhos, Jonathan, Kogkitsidou, Eleni, Gambette, Philippe, Sagot, Benoît et Gabay, Simon. « Automatic Normalisation of Early Modern French ». In : *LREC 2022 - 13th Language Resources and Evaluation Conference*. European Language Resources Association. Marseille, France, 2022, p. 3354-3366. URL : https://inria.hal.science/hal-03540226.

[2] Burns, Patrick J. « LatinCy : Synthetic Trained Pipelines for Latin NLP ». 2023. arXiv : 2305.04365 [cs.CL]. URL : https://arxiv.org/abs/2305.04365.

[3] Camps, Jean-Baptiste, Clérice, Thibault, Kanaoka, Naomi, Pinche, Ariane, Duval, Frédéric et Ing, Lucence. « Corpus and Models for Lemmatisation and POS-tagging of Old French ». Document de travail. 2021. URL : https://shs.hal.science/halshs-03353125.

[4] Chagué, Alix et Clérice, Thibault. « "I'm Here to Fight for Ground Truth" : HTR-United, a Solution Towards a Common for HTR Training Data ». In : *Digital Humanities 2023 : Collaboration as Opportunity*. Alliance of Digital Humanities Organizations and University of Graz. Graz, Autriche, 2023. URL : https://inria.hal.science/hal-04094233.

[5] Clérice, Thibault. *Détection d'isotopies par apprentissage profond : l'exemple de la sexualité en latin classique et tardif*. Thèse de doctorat dirigée par Nicolas, Christian Lettres et civilisations antiques Lyon 2022. Thèse de doct. Lyon : Université de Lyon, 2022. URL : http://www.theses.fr/2022LYSE3007/document.

[6] Clérice, Thibault. « Pie Extended, an extension for Pie with pre-processing and post-processing ». 2020. DOI : 10.5281/zenodo.3883589.

[7] Clérice, Thibault, Janes, Juliette, Scheithauer, Hugo, Bénière, Sarah, Cafiero, Florian, Romary, Laurent, Gabay, Simon et Sagot, Benoît. « Diachronic Document Dataset for Semantic Layout Analysis ». Document de travail. 2024. URL : https://hal.science/hal-04784161.

[8] Clérice, Thibault, Jolivet, Vincent et Pilla, Julien. « Building Infrastructure for Annotating Medieval, Classical and Pre-Orthographic Languages : The Pyrrha Ecosystem ». In : *Digital Humanities 2022 (DH2022)*. Tokyo, Japon : Alliance of Digital Humanities Organizations, 2022. URL : https://hal.science/hal-03606756.

[9] Clérice, Thibault, Vlachou-Efstathiou, Malamatenia et Chagué, Alix. « CREMMA Medii Aevi : Literary Manuscript Text Recognition in Latin ». In : *Journal of Open Humanities Data 9* (2023), p. 4. DOI : 10.5334/johd.97.

[10] Clérice, Thibault et al. « CATMuS Medieval : A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond ». In : *2024 International Conference on Document Analysis and Recognition (ICDAR)*. Athène, Grèce, 2024. URL : https://inria.hal.science/hal-04453952.

[11] Daneau, Lambert. *In D. Pauli priorem Epistolam ad Timotheum commentarius*. BGE Cti 1753 BGE S 22877. Geneva : apud Eustathium Vignon, 1577. DOI : 10.3931/e-rara-6338.

[12] Forcellini, Aegidius, Furlanetto, Iosephus, Corradini, Franciscus et Perin, Iosephus. *Lexicon Totius Latinitatis*. Padoue : Typis Seminarii, 1940.

[13] Gabay, Simon. « FONDUE-FR-PRINT-17, - Transcriptions of French 17th c. prints ». 2024. DOI : 10.5281/zenodo.11526040.

[14] Gabay, Simon. « FONDUE-LA-PRINT-16 - Transcriptions of French 16th c. prints ». 2024. DOI : 10.5281/zenodo.11526149.

[15] GABAY, Simon et BARRAULT, Loïc. « Traduction automatique pour la normalisation du français du XVIIe siècle ». French. In : *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, sous la dir. de Christophe BENZITOUN, Chloé BRAUD, Laurine HUBER, David LANGLOIS, Slim OUNI, Sylvain POGODALLA et Stéphane SCHNEIDER. Nancy, France : ATALA et AFCP, 2020, p. 213-222. URL : https://aclanthology.org/2020.jeptalnrecital-taln.20.

[16] GABAY, Simon, CHAMPENOIS, Robin, KUENZLI, Pierre, FALCONE, Jean-Luc et CHARPILLOZ, Christophe. « Formes Numérisées et Détection Unifiée des Écritures (FoNDUE) ». 2021. URL : https://fondue.unige.ch.

[17] GABAY, Simon et CLÉRICE, Thibault. « The Birth of French Orthography : A Computational Analysis of French Spelling Systems in Diachrony ». In : *CHR2024 – Computational Humanities Research Conference*. Aahrus, Danemark, 2024. URL : https://inria.hal.science/hal-04704549.

[18] GABAY, Simon, CLÉRICE, Thibault, CAMPS, Jean-Baptiste, TANGUY, Jean-Baptiste et GILLE-LEVENSON, Matthias. « Standardizing Linguistic Data : Methods and Tools for Annotating (Pre-Orthographic) French ». In : *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC '20)*. Hammamet, Tunisie, 2020. DOI : 10.1145/3423603.3423996. URL : https://hal.science/hal-03018381.

[19] GABAY, Simon, GOY, Floriane et JEANNOT-TIROLE, Marie. « FONDUE-LA-PRINT-16 - Transcriptions of Latin 16th c. prints ». 2024. DOI : 10.5281/zenodo.11526159.

[20] GABAY, Simon, PINCHE, Ariane, CHRISTENSEN, Kelly et CAMPS, Jean-Baptiste. « SegmOnto : A Controlled Vocabulary to Describe and Process Digital Facsimiles ». In : *Journal of Data Mining and Digital Humanities* (2024). DOI : 10.46298/jdmdh.12689.

[21] GABAY, Simon et al. « Reconnaissance des écritures dans les imprimés ». In : *Humanistica 2024*. OCR. Association francophone des humanités numériques. Meknès, Maroc, 2024. URL : https://hal.science/hal-04557457.

[22] GOY, Floriane. « gallicorpora_ajuste ». Version v.1.2.0. 2024. DOI : 10.5281/zenodo.19218113. URL : https://doi.org/10.5281/zenodo.19218113.

[23] GOY, Floriane. « Layout-16th-Print-Lat ». Version v1.0.0. 2026. DOI : 10.5281/zenodo.18492102.

[24] HUMEAU, Maxime, GABAY, Simon et PINCHE, Ariane. « SegmOnto ». Version boscaiola. 2024. DOI : 10.5281/zenodo.10602197.

[25] JOHNSON, Kyle P., BURNS, Patrick J., STEWART, John, COOK, Todd, BESNIER, Clément et MATTINGLY, William J. B. « The Classical Language Toolkit : An NLP Framework for Pre-Modern Languages ». In : *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations*, sous la dir. d'Heng JI, Jong C. PARK et Rui XIA. Online : Association for Computational Linguistics, 2021, p. 20-29. DOI : 10.18653/v1/2021.acl-demo.3.

[26] KIESSLING, Benjamin. « Kraken - an Universal Text Recognizer for the Humanities ». en. In : *Digital Humanities Conference 2019 - DH2019*. Alliance of Digital Humanities Organizations and University of Graz. Utrecht, Pays-Bas, 2019. DOI : 10.34894/Z9G2EX.

[27] KIESSLING, Benjamin, TISSOT, Robin, STOKES, Peter et STÖKL BEN EZRA, Daniel. « eScriptorium : An Open Source Platform for Historical Document Analysis ». In : *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). T. 2. 2019. DOI : 10.1109/ICDARW.2019.10032.

[28] MORETTI, Franco. *Distant Reading*. Londres/New-York : Verso, 2013.

[29] PINCHE, Ariane et GABAY, Simon. « Gallicorpora+ ». 2022. DOI : 10.5281/zenodo.7410529.

[30] PINCHE, Ariane et al. « CATMuS-Medieval : Consistent Approaches to Transcribing ManuScripts ». In : *Digital Humanities - DH2024*. Online : hal-04346939. Alliance of Digital Humanities Organizations and University of Graz. Washington D.C., États-Unis, 2024. URL : https://inria.hal.science/hal-04346939.

[31] PINCHE et GABAY. « Gallicorpora+ ». 2022. DOI : 10.5281/zenodo.7410529.

[32] QI, Peng, ZHANG, Yuhao, ZHANG, Yuhui, BOLTON, Jason et MANNING, Christopher D. « Stanza : A Python Natural Language Processing Toolkit for Many Human Languages ». 2020. arXiv : 2003.07082 [cs.CL]. URL : https://arxiv.org/abs/2003.07082.

[33] SCHÖCH, Christof. « Topic Modeling Genre : An Exploration of French Classical and Enlightenment Drama ». In : *Digital Humanities Quarterly 11* (2 2017). DOI : 10.63744/qa2r7vpsu35k.

[34] SOLFRINI, Sonia, GABAY, Simon, HUMEAU, Maxime, PINCHE, Ariane, BEAULNES, Pierre-Olivier, OLIVEIRA, Aurélia Marques, GROSS, Geneviève et CAMILLOCCI, Daniela Solfaroli. « Océriser les imprimés du XVIe siècle en langue française ». In : *Humanistica 2024*. OCR. Association francophone des humanités numériques. Meknès, Maroc, 2024. URL : https://hal.science/hal-04555002.

[35] SPEYER, Miriam. « Les dieux écrivent-ils en italiques ? Typographie et mise en livre de pièces en vers et en prose ». In : *L'Habillage du livre et du texte aux XVIIe et XVIIIe siècles*. T. 9. (Book Practices & Textual Itineraries). PUN, Éditions Universitaires de Lorraine, 2019, p. 79-92. URL : https://normandie-univ.hal.science/hal-02184237.

[36] WALLACE, William. « An Index of Greek Ligatures and Contractions ». In : *The Journal of Hellenic Studies 43*, no. Part 2 (1923), p. 183-193. URL : https://www.jstor.org/stable/625810.

[37] ZAHND, Ueli. « Lambert Daneau kommentiert Petrus Lombardus – Eine reformierte Auseinandersetzung mit einem Basistext mittelalterlicher Scholastik ». In : *Die Reformation und ihr Mittelalter*, sous la dir. de Günter FRANK et Volker LEPPIN. Stuttgart : Frommann-Holzboog, 2015, p. 263-282. DOI : 10.5771/9783772830853-263.

[38] ZAHND, Ueli, KRAUTER, Stefan, COLOMBO, Matteo, GOY, Floriane, MANIG, Benjamin et SCHÜRMANN, Noemi. « 16th Century Exegesis of Paul ». Grant number SNFS : 207696. Genève, Zürich, 2023. URL : https://www.unige.ch/ihr/fr/accueil/exegese-paulinienne.

A Nommage des zones

— `RunningTitleZone` pour le titre courant en haut de page ;
— `MarginTextZone-Notes` pour les commentaires marginaux ;
— `NumberingZone` pour la numérotation (pagination ou foliotation) ;
— `QuireMarksZone` pour l'organisation des cahiers ;

— `GraphicZone` pour toutes les décorations (illustrations, ornementation…) ;
— `DropCapitalZone` pour les lettrines en début de partie ;
— `MainZone-P` pour un paragraphe ;
— `MainZone-Continued` pour la suite d'un paragraphe interrompu ;
— `MainZone-Head` pour le titre intermédiaire et les versets bibliques, lorsqu'ils sont mis en valeur par la mise en page

Les lignes ont été classées en deux catégories :

— `HeadingLine` continent tout les éléments de titre, sous-titre, etc. ;
— `DefaultLine` décrit la ligne par défaut.