Wauchier, Is That You? A Multi-Manuscript Authorship Analysis of Saint Lambert's Life

Thibault Clérice[1], and Ariane Pinche[2]

[1] Inria, Paris, France
[2] CIHAM UMR 5648, CNRS, Lyon, France

Abstract

We investigate the possible authorship of the *Vie de saint Lambert*, an anonymous Old French hagiographic text. Building on previous research that noted a stylistic proximity between this text and the *Seint Confessor* attributed to Wauchier de Denain, we expand the inquiry across ten manuscripts. Alongside our computational analysis, we provide a brief study of the manuscript tradition and the transmission of the Vita of Saint Lambert within the broader literary genre of hagiography. We work directly with automatic transcriptions from medieval manuscripts, enhancing the tooling for post-processing space errors and lemmatization. We employ the recent Bootstrap Distance Impostors (BDI) approach to improve precision in authorship verification tasks, and propose a methodology for addressing the challenges of massively anonymous corpora in this context. Our results confirm a tendency to attribute the *Vie de saint Lambert* to Wauchier de Denain, though the stylistic alignment is less marked than in texts firmly attributed to him.

Keywords: Automatic text recognition, Autorship attribution, Compilation, Hagiography, Medieval Manuscripts

Introduction

The notion of authorship in medieval French literature is often elusive. While a handful of names – such as Chrétien de Troyes or Marie de France – have come down to us, many works remain unattributed. Some other identifications are even challenging, as the identification of the author(s) of the *Roman de la Rose* or of the *Roman de Renart*. This anonymity is even more current in hagiographic literature, where most texts circulate without attribution, although a few names stand out, such as Jean de Vignay, Wace, or Wauchier de Denain. Hagiographic manuscripts from the 13th century are compilations that combine mostly anonymous Lives of saints with a minority of texts attributed to known authors. Wauchier de Denain, a prolific 13th-century writer, is known to be the author of two coherent collections of saints' lives: *Li Seint Confessor* [49] and the *Vies des saints Peres d'Egipte* [48].

During the 2019 *Digital Humanities Conference*, Camps, Clérice, and Pinche [7] demonstrated that applying stylometry directly to Automatic Text Recognition (ATR) output from a medieval manuscript could yield promising results. Such a pipeline, which operates on automatically produced and inherently imperfect transcriptions, requires working with noisy data: unnormalized texts, spelling variation, lack of modern punctuation, and inconsistent tokenization. Their case study, based on a single manuscript (BnF, fr. 412), confirmed a longstanding hypothesis by Paul Meyer regarding the structure of French hagiographic compilations, *i.e.* that they were assembled from pre-existing thematic units. The analysis also revealed an unexplained proximity between the Lives from Wauchier de Denain's *Seint Confessor* and an anonymous Life of Saint Lambert.

Thibault Clérice, and Ariane Pinche. "Wauchier, Is That You? A Multi-Manuscript Authorship Analysis of Saint Lambert's Life." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 132–148. https://doi.org/10.63744/QsBV0XYj8wRC.

Given that most of the Lives of Saints in Old French are translations of Latin sources, attributing a text to a known author could provide valuable insights into the context of its creation, its intended audience, and the reasons behind the emergence of vernacular prose hagiography in the early 13th century.

Building on this preliminary observation and recent improvements in both preprocessing (ATR specifically) and stylometry, specifically with Nagy's recent extension of the impostors approach [35], the present study aims to further investigate the authorship of the *Vie de saint Lambert* by combining ATR and stylometric analysis across a broader corpus of manuscript witnesses. Rather than relying on edited or normalized texts, which may introduce editorial bias or alter original linguistic features, we work directly from automatic transcriptions, embracing the challenges of noisy, uncorrected data and enabling a more direct computational engagement with the materiality of medieval manuscripts. While the central question remains whether the *Vie de saint Lambert* can be attributed to Wauchier de Denain, our objectives are twofold: first, to strengthen the hypothesis by testing it across ten manuscript witnesses (See Table 1); and second, to refine the entire pipeline of data acquisition and preparation prior to stylometric analysis. By doing so, we aim not only to clarify a specific case of medieval authorship, but also to propose a reproducible methodology applicable to other unedited or partially edited medieval texts.

The outcomes of this paper are:

- a corpus of ground-truth annotations used to benchmark our tools in ATR, layout analysis, word segmentation and abbreviation resolution;[1]
- a new corpus of ten manuscripts, split by text according to Jonas database[2], with full-text recognition and various layers of annotation, including lemmatization and POS-tagging;[3]
- an adaptation of the previous work of Clérice [10] to resegment words in ATR outputs;[4]
- tooling to use a YOLO [25] object detection model to do layout analysis and produce ALTO XML, a format compatible with the Kraken ATR library;[5]
- an analysis of Wauchier de Denain's relationship with the *Vie de Saint Lambert*.

After a survey of related works in the realms of ATR and related technologies applied to medieval documents as well as of stylometry and, specifically, in a noisy context, we will introduce more the current state of the art regarding both French hagiographic collections and the subject of our paper, Wauchier de Denain and the Life of Saint Lambert. We will then explain our experiment, followed by a result analysis and a conclusion.

1 Related work

1.1 Handwritten Text Recognition for medieval manuscripts

ATR for medieval manuscripts has experienced significant advancements in recent years, driven by both improvements in model architectures and the availability of large, high-quality datasets. On the one hand, state-of-the-art neural network architectures have become widely accessible through off-the-shelf tools and software. For example, convolutional recurrent neural networks (RCNNs)

[1] The ground-truth can be found at `https://github.com/PonteIneptique/st-lambert/`.

[2] The Jonas database is a repertory of French and Occitan literary texts from the Middle Ages and early modern times. This resource is produced and distributed by the IRHT for research consultation, `https://jonas.irht.cnrs.fr`.

[3] The pipeline for building and analysing the data, and the data are available at `https://github.com/PonteIneptique/st-lambert-comput`.

[4] `https://github.com/PonteIneptique/boudams2`.

[5] `https://github.com/PonteIneptique/yolalto`.

are implemented effectively in frameworks such as Kraken [27] and PyLaia [45], which allow researchers to train and deploy ATR models with relative ease. More sophisticated architectures, including transformer-based models like TrOCR [1] and Party [26], push the performance envelope further, albeit requiring more computational resources and expertise to train and fine-tune.

On the other hand, recent efforts have substantially expanded the corpus of annotated medieval manuscript data available for training and evaluation. Notably, the Tridris project [1] and the CATMuS corpus [12] have contributed hundreds of thousands of labeled lines, providing a rich resource for supervised learning approaches. Since late 2024, CATMuS has also released segmentation datasets that facilitate the training of layout analysis models specifically tailored to the complexities of medieval manuscript page structures.[6] These datasets cover a wide variety of scripts and layouts from the Middle Ages, enabling the development of robust segmentation and recognition pipelines.

The combination of these well-curated datasets and advanced model architectures has resulted in significant improvements in recognition accuracy. For instance, the CATMuS Medieval 1.6.0 model [39] achieves character accuracy over 95% in optimal conditions, which represents a major milestone given the intrinsic challenges of medieval handwriting, such as irregular letterforms, variable abbreviations, and degraded manuscript quality. From CATMuS 1.6.0, [9] has shown that a few hundred lines could yield significant character accuracy improvement, up to 30%.

1.2 Stylometry methods

The landscape of stylometry, specifically in authorship verification and authorship attribution, has seen limited advancement since our work on Wauchier first presented in 2019 [7]. Despite increasing interest, improvements to existing approaches have been at best incremental. The feature sets in use remain largely the same: most frequent words [18; 41], function words, affixes [7], and POS n-grams [5]. Less common but still noteworthy features include rhymes [5], metrical patterns [36], and syntactic annotations [20], although the latter are often constrained by annotation costs and limited tool availability. Noteworthy is the work of Eder regarding boosting the support of features [17].

In terms of methodology, both supervised and unsupervised approaches continue to favor interpretability. Models such as SVMs and hierarchical clustering, using distance metrics like Burrows' Delta [4], Eder's Delta [19], Manhattan, or Cosine distance, remain prevalent. This is consistent with the emphasis on explainability in the computational and digital humanities. In authorship verification, methods like General Imposters (GI) [28], introduced to the digital humanities by [24], have expanded the available toolkit. The General Impostors method compares a target text with candidate authors and impostors (authors or texts assumed not to be the true author) across multiple runs. In each run, a randomly selected subset of features is used, and the candidate most frequently identified as closest to the target text—according to a chosen distance metric—is considered the author. Nagy's Bootstrap Distance Impostors (BDI, [35]) builds on GI by emphasizing precision over recall. BDI relativizes the original impostor percentage score (% of run where the attribution is made with the candidate) using a distance-based metric, enabling density estimation and interpretation as a statistical likelihood, as proposed by Nagy. Less interpretable yet reusing the same standard features (*i.e.* full text, but rather POS, affixes, etc.), [11] provides a deep learning approach with feature embedding comparison in a siamese network aimed at separating positive and negative pairs of texts.

Stylometry on large collections of anonymous texts remains a major challenge, as most authorship verification methods rely on predefined positive or negative pairings. In the context of medieval corpora, feature-based approaches remain rare. [8] applies such techniques to a small,

[6] `https://huggingface.co/biglam/medieval-manuscript-yolov11`

controlled candidate set. [2] presents a direct reuse of BDI, while [47] employs character n-grams for scribal attribution and dating in a parallel line of inquiry. [22] evaluates precision and recall for specific hyperparameters (e.g., number of most frequent words) for early modern Spanish, which appears to be edited as such and to bear the marks of editorial work. Overall, stylometric work on vernacular romance texts, specifically in their raw form, from the medieval period remains limited, in part due to the specific challenges posed by unfixed spelling, diachronic linguistic change, and transmission variability.

2 Saint Lambert's and its context

2.1 French hagiographic collection

As revealed by Paul Meyer [34], French hagiographic manuscripts often result from the accumulation of thematic collections of saints' lives. Meyer identified three major families of texts, which he labelled A, B, and C, each corresponding to a specific thematic unit: family A focuses on the Apostles, B on the Martyrs, and C on the Confessors. These units circulated independently or in combination, with manuscripts containing A, A+B, or the full A+B+C set, as exemplified by BnF fr. 412. While this thematic structure seemed to dominate French hagiographic compilations in the 13th century, their organisation gradually shifted over time. Later manuscripts increasingly adopt a structure aligned with the liturgical calendar, following the model of the *Legenda Aurea*.[7]

Understanding these structural dynamics is crucial not only for the study of compilation practices but also to shed light on the development of French vernacular hagiography and the role it played within medieval devotional and literary culture.

In particular, it may help us understand how the earliest compilations were constructed. As Guy Philippart has argued for Latin hagiographic collections, early compilations may have drawn on pre-existing thematic or authorial booklets [37]. Investigating similar principles applied to Old French prose hagiography can shed light on how hagiographic texts circulated and were grouped in manuscript contexts.[8]

2.2 Saint Lambert's Life and Wauchier de Denain

As an initial step, we focus on the *Vie de saint Lambert*, a text classified within Family B (Martyrs), which was previously associated with the Lives of the *Seint Confessor* in manuscript BnF fr. 412 through the research lead by Camps, Clérice, and Pinche. This version is a long French prose *Vita* of the saint, based on the earliest Latin traditions.

Lambert was bishop of Tongres-Maastricht during the reign of Childeric II and later under Pepin of Herstal, father of Charles Martel. He was assassinated between 696 and 705, and his relics were transferred to Liège, where his cult remains highly popular in Belgium to this day [29]. The Latin Lives of Saint Lambert were widely disseminated in medieval manuscripts, and a short version was incorporated into the *Legenda Aurea* [30].[9]

Two major versions of the Latin Vita have come down to us. The earliest version, originally written by an anonymous author and later rewritten in more "elegant" Latin by Etienne, bishop of Liège, portrays Lambert's death as the outcome of a local conflict between aristocratic families over

[7] The *Legenda Aurea* (Golden Legend) is a widely circulated Latin collection of saints' lives compiled in the 13th century by Jacobus de Voragine, a Dominican friar and later Archbishop of Genoa. Organized according to the liturgical calendar, the work includes over 150 hagiographic narratives and became a standard reference for preachers and compilers of vernacular collections throughout medieval Europe. Its immense popularity led to numerous translations and adaptations, including into Old French, and it often served as a model for the structure and content of later hagiographic compilations.

[8] Initial investigations have been made about the legendier Picard by Labie-Leurquin [31].

[9] The link to the list of the different versions are available in the base *Légendiers latins*: "Lambertus04 = Lambertus ep. Traiectensis" (permalink: `https://legendiers-latins.irht.cnrs.fr/37664`), accessed on 04/07/2025.

control of episcopal property. A later version – popularized through the *Legenda Aurea* and widely transmitted by authors such as Adon of Vienne, Hucbald of Saint-Amand, Sigebert of Gembloux, and Nicolas of Saint-Denis – frames his martyrdom as a punishment for condemning the illicit relationship between Pepin and his concubine Alpaïda.

Both versions were translated into the French vernacular hagiographic manuscripts. According to the Jonas database, 73 French witnesses of the *Vie de saint Lambert* have been recorded. Among them, two principal versions of the text can be distinguished: one, identified as *Jonas 3058*, consists of 30 witnesses and corresponds to the 14th-century translation of the *Legenda Aurea* by Jean de Vignay. The second, *Jonas 1958*, has 20 witnesses – six of which date from the 13th century – and remains anonymous. This latter, older version is the one found in BnF fr. 412 and the one previously identified as potentially close in style to Wauchier de Denain's *Seint Confessor*.

Chronologically, the hypothesis is plausible. Wauchier de Denain was active in the early 13th century. Furthermore, both of his major hagiographic collections, *Li Seint Confessor* and *Les Vies des saints peres d'Egipte*, consist of French prose translations of renowned Latin *Vitae*: the Life of saint Martin (from Sulpicius Severus), Life of saint Nicolas, and Life of saint Benoit (from Gregory the Great) [49], suggesting that Wauchier was a well-educated cleric with access to reputable sources. As Paul Meyer noted, Wauchier's Lives are among the earliest vernacular prose versions of these texts and often present extensive and detailed narratives – much like the long version of the *Vie de saint Lambert* (*Jonas 1958*).

Other converging elements strengthen the attribution hypothesis. Wauchier de Denain's works are associated with the court of Flanders, and writting about saints linked to that region – such as Lambert – would be consistent with his known affiliations. Significantly, the Latin Vita that may have served as the source for Wauchier's version is preserved in the 12th-century manuscript Douai 856, which belonged to the library of the Abbey of Marchienne.[10] This abbey was part of a network of book exchanges during the 12th and 13th centuries that included the armarium of Saint-Amand [43]. The latter is known to have contained several of the Latin sources used by Wauchier de Denain [16], as attested by surviving inventories such as the Index Major [14]. Moreover, it is well documented that Etienne of Liege[11] and Hucbald of Saint-Amand[12] were in contact and corresponded about their writing projects [15]. This interconnection suggests that the Latin source material was accessible to Wauchier de Denain or someone working in the same area.

Thematically, the *Vie de saint Lambert* aligns closely with the *Seint Confessor* corpus: the protagonist is a powerful bishop and spiritual advisor to rulers (as are Martin with the Emperor Julian, Benoit with Totila, or Martial with Duke Stephen). Furthermore, the theme of obedience – a central moral value in Wauchier's works – is also prominent in the text, notably in an episode where Lambert nearly dies of cold after obeying an abbot's order (similar stories appear in the *Dialogues sur les vertus de saint Martin* or *Vie de saint Benoit*). However, thematic similarities alone are insufficient. Certain stylistic features characteristic of Wauchier are absent: the *Vie de saint Lambert* does not contain Latin insertions translated into French, nor poetic interpolations in prose, which are often found in Wauchier's corpus. Moreover, no authorial signature or dedicatory address has been identified in any of the digitized witnesses we consulted. These limitations led us to turn to stylometric analysis in the hope of uncovering new evidence to support or refute this potential attribution.

[10] See description: `https://ccfr.bnf.fr/portailccfr/jsp/index_view_direct_anonymous.jsp?record=eadcgm:EADC:D06A14123`.

[11] Etienne, bishop of Liège from 901 to 920, is known for reworking the earliest and most historically grounded version of the Vita of Saint Lambert.

[12] Hucbald of Saint-Amand (d. after 930) was a monk and hagiographer affiliated with the abbey of Saint-Amand. He is the author of a Vita of Saint Lambert that follows the narrative structure of the second Latin version of the saint's life.

3 Experiments

3.1 Data input and preprocessing

Manuscripts We extend the experiment beyond the original study [7] by incorporating ten manuscripts representing various families and witnesses of French legendiers according to Paul Meyer. The selected manuscripts are listed in Table 1. The corpus used in this study comprises several manuscripts belonging to different legendary groupings, each offering varying degrees of proximity to Wauchier de Denain's corpus. Among these manuscripts, BnF fr. 412 was already analyzed in [7].

City	Repository	Shelfmark	Date	Family	Provenance	Nb of W. Texts	Nb of S. Conf. Texts
Paris	Bibliothèque Sainte-Geneviève	ms, 588	1296	B	île de france	0	0
Paris	Bibliothèque nationale de France	ms, fr. 00412	1285	C	Hainaut	9	9
Paris	Bibliothèque nationale de France	ms, fr. 00411	1301-1400	C	Northern France	9	9
Paris	Bibliothèque nationale de France	ms, fr. 17229	1201-1300	D	unknown	1	1
Paris	Bibliothèque nationale de France	ms, fr. 06447	1275	D1	Northern France	3	2
Chantilly	Bibliothèque du Château	0734 (0456)	1312	E	Paris	7	7
Basel	Universitätsbibliothek	Com. Lat. 102	1320-1330	E	Paris	7	7
Paris	Bibliothèque nationale de France	ms, fr. 23117	1301-1400	F	unknown	7	7
Paris	Bibliothèque nationale de France	ms, fr. 00185	1301-1400	G	unknown	10	7
Paris	Bibliothèque nationale de France	NAF 23686	1250	L	unknown	7	6

Table 1: List of the manuscripts used during the experiments

BSG 588, a representative of Legendier Group B, contains no works attributed to Wauchier de Denain, as it preserves only texts related to the theme of martyrdom. Therefore, confessors, the central figures of Wauchier's collections, are absent from this compilation.

BnF fr. 412 and BnF fr. 411, both representatives of Group C, preserve the most complete collections of Lives attributed to Wauchier de Denain. This family is the only one to transmit the full and ordered series of the *Seint Confessor*, including the *Dialogues sur les Vertus de saint Martin*. These witnesses, with the manuscript London, British Library, Royal 20.D.VI,[13] are considered the closest to the archetype of *Li Seint Confessor*, especially in their preservation of verse passages, dedicatory addresses, and authorial signatures. We can also observe that the *Vie de saint Lambert* is most frequently found in manuscripts that include the *Seint Confessor*, rather than those containing the *Vies des saints peres d'Egipte*.

BnF fr. 17229 and BnF fr. 6447 are the last surviving witnesses of Compilation D. However, fr. 6447 presents substantial textual variation in the Life of saint Martin. This version is classified in Group M of the stemma of the *Seint Confessor* and shows affinities with the variant readings found in Legendary Groups G and L. [49] This manuscript is the only one of the corpus to contain a Life of sainte Marthe written by Wauchier the Denain.

Manuscripts from Group E (Chantilly and Basel) are also closely related, both in terms of composition and textual affiliation with the *Seint Confessor*. Their production history suggests proximity as well. Both appear to originate from Parisian workshops: Chantilly's in 1312, and Basel's between 1320 and 1330. The miniatures in the Chantilly manuscript have been stylistically linked to Richard de Verdun, who may be the same artist as the illuminator known as the Maître de Papeleu, although this identification remains debated [13; 42; 44]. This connection is particularly noteworthy, as the Maître de Papeleu has also been identified as the illuminator of the Basel manuscript [23]. These shared stylistic features may indicate that both manuscripts were produced within the same or closely connected ateliers.

In Groups G and L, at least for the Lives of Saint Martin and Saint Nicolas attributed to Wauchier de Denain, a second redaction has been identified.[14] This later version is character-

[13] Unfortunately, no digitization of the manuscript is now available.

[14] This observation was made by J. J. Thompson for the Life of Saint Nicolas [16] and by Pinche for the Life of Saint Martin [49].

ized by the omission of verse passages and a partial modernization of the language, particularly in the use of function words [49]. Those manuscripts are also the only one of our corpus to mix Lifes from *Li Seint Confessor* and *Vies des saints peres d'Egipte*. It should be noted that BnF NAF 23686 has been mutilated and contains numerous missing folios, resulting in partial losses of text throughout the manuscript.

All manuscript images were obtained from their respective repositories and processed through a semi-automated pipeline, described in the following section.

Text Recognition Our pipeline builds upon the CATMuS Medieval datasets [12], including both the textual and segmentation components. We employ a YOLOv11 model trained by W. Mattingly [33], capable of detecting both zones and lines, in conjunction with CATMuS Medieval 1.6.0 [40], a Kraken-compatible model.

To enhance the YOLOv11 predictions, we developed a custom Python adaptation that integrates the region-line binding and line-reordering heuristics from eScriptorium, along with a centerline generation procedure–derived from the bounding box centroid to meet Kraken's requirements. Within eScriptorium, segmentation results are manually reviewed to correct major errors, and incipit lines are annotated in preparation for later extraction of distinct literary units.

We corrected a selection of pages from four manuscripts (Chantilly 734, CL 102, BnF Fr. 411, 6447, and 23117), resulting in a dataset of 1,017 lines for training and validation (excluding 23117), and 653 lines for testing. This data was used to fine-tune the generic CATMuS model, producing a new evaluation model (Table 2). Average character accuracy (CA) improves overall (+1.73%), one manuscript showed a significant performance decrease (-0.83%), but our only manuscript in test showed more than 1% of improvement. The fine-tune is essential to address the reuse of the YOLO segmentation; prior research has shown that Kraken and similar models are sensitive to segmentation variation between training set and inference situations [3; 38].

Model	Lines	Fr. 411	Chant., 734	CL 102	Fr. 6447	Fr. 23117, H1	Fr. 23117, H2	Micro-average
Catmus 1.6.0	190,000	95.92	**93.05**	97.78	96.74	92.56	94.12	93.21
– Finetuned	+ 1017	**97.00**	92.22	**98.07**	**97.04**	**94.37**	**95.53**	**94.94**

Table 2: Character Accuracy for CATMuS 1.6.0 model results and its fine-tune over our 5 test manuscripts. H stands for Hand.

Text Post-Processing After identifying paratextual elements via Segmonto layout annotation, we extract individual works (literary units), such as the Life of saint Lambert or saint Martin, from each manuscript. Abbreviation patterns are scribe-specific, and they pose a blocking issue when comparing multiple manuscripts to each other, as each scribe may have a different abbreviation rate and practice. Given the original results of [7], we hypothesize that abbreviation noise has limited impact within a manuscript, especially if it has only one scribe.[15] As our approach involves only intra-manuscript comparison, and not inter-manuscript comparison, we depart from our 2021 work and do not resolve abbreviations.

In medieval ATR, space is often responsible for a third of the errors, as spacing is fluctuating in manuscripts. We adapted the Boudams retokenizer [10] to address the situation, to retokenize the output of the ATR. Unlike the 2019 analysis of BnF, fr. 412, our implementation preserves original spacing and introduces a third character class to remove redundant spaces.[16] Missing and superfluous spaces respectively account for 12% and 9.3% of tokenization errors in our fine-tuned model, not including hyphenation issues.

[15] Most manuscripts have a single hand, and others two, e.g. BnF fr. 185 (change of hand fol. 274) and BnF fr. 23117 (fol. 228).

[16] In our Boudams version, characters are categorized into: (1) Keep Class–for valid spaces and in-word characters, (2) Word-Ending Class–for inserting missing spaces, and (3) Extraneous Class–for removing superfluous spaces.

In addition to this adaptation, we trained the model on the full BFM corpus [21], augmented with synthetic spacing and abbreviation, as well as synthetic ATR errors after the spacing and abbreviation step. Evaluation against manual annotations is reported in Table 3, using raw ATR output (not ground truth) as input. Overall, not only does adding a class for removing spaces improve the results, but the model enhances spacing by up to 69%, with only one case of (light) degradation for our overall best model configuration (CNN 7).

	Bnf Naf 23686	Bnf fr. 10128	Bnf fr. 23112	Bnf fr. 23117	Bnf fr. 24429	Sainte Genevieve 588
Boudams V1 (CNN 5)	29.96	**72.53**	-3.24	-63.39	-121.51	-15.88
Ours (CNN 5)	54.05	69.55	23.01	3.47	-36.19	-24.86
Ours (CNN 7)	**55.47**	66.74	15.68	**6.18**	0.14	-4.21
Ours (CNN 9)	46.71	69.72	15.68	-8.90	**5.20**	-24.86
Ours (CNN 11)	40.63	65.27	**23.11**	3.47	-25.89	-10.01
Ours (RNN 1)	-0.48	69.55	-56.84	-41.21	-72.74	-45.89
Ours (RNN 2)	46.42	69.63	23.01	-21.79	-21.05	**0.31**

Table 3: Character accuracy improvement on inference text. Negative values indicate a degradation of content: -121% means that errors are more than doubled with the model. Positive numbers indicate that spacing was improved: 100% would mean that no more spacing errors are present in the text. Overall, space correction (insertion+deletion) outperforms space insertion. Model variants are denoted as follows: (RNN X) refers to X layers of bidirectional GRUs, while (CNN X) indicates a sequence of 10 convolutional layers with kernel size X.

We then fine-tuned a model for Old French part-of-speech tagging and lemmatization using the latest PaPie version, including O. Nedey implementation of vocabulary extension for fine-tuning [32]. Training was performed for five epochs on the OF3C training set [6], incorporating the same synthetic abbreviation and spacing rules as for the retokenizer. While performance metrics in Table 4 show a general drop in lemmatization and tagging accuracy, this decline is smaller when using the model fine-tuned on synthetic abbreviation data. As expected, lemmatization largely breaks the original clustering of data per manuscript by removing at least period or scribal noise (see 1).

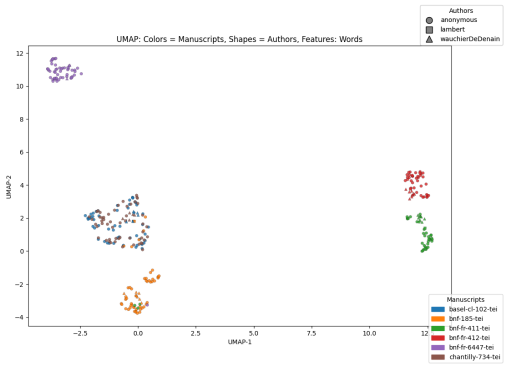
			POS		Lemmatization
		Accuracy	Precision	Recall	Accuracy
Original Test Set	OF3C Model	94.82	73.19	73.2	93.65
Corrupted Test Set	OF3C Model	87.74	61.69	60.95	83.02
	– Finetuned	**89.66**	**64.32**	**64.56**	**86.65**

Table 4: Lemmatizer results using PaPie, in percent.

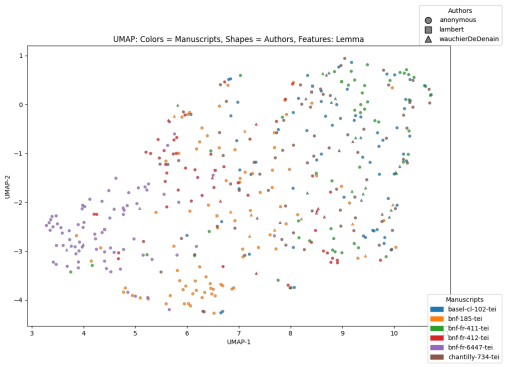
The resulting corpus shows a median length around 3000 tokens, with two manuscripts (BSG 588 and Basel CL 102) having the smallest texts of the whole corpus, which is coherent with the Legendier's classification of Meyer (see Table 2a).

3.2 Experimental set-up

Like most authorship verification system, approaches such as the Bootstrap Distance Impostor (BDI) from Nagy [35] aim at precision over recall, where false positives are much more problematic than false negatives. Given this situation, we design an experiment around (1) each manuscript and,



(a) UMAP of each individual work using the global 200 most frequent words' relative frequency.

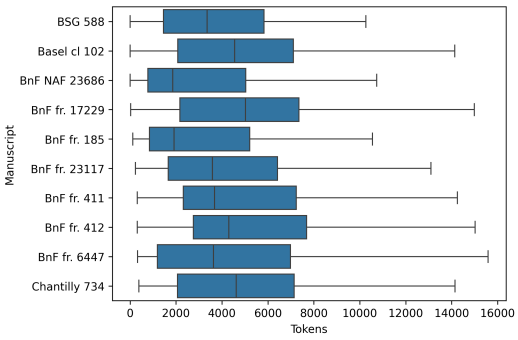


(b) UMAP of each individual work using the global 200 most frequent lemmas' relative frequency.

Figure 1: UMAP representations of works accross manuscripts, using two different set of features (words vs. lemma). On the left, we see each manuscript clustered, with manuscript in the same legendier branch being closer to each other. On the right, the signal of legendiers branches as well as the manuscript is mostly gone, except for the manuscript Fr. 185 (Purple) whose lives vary heavily from the other branches.

Manuscript	Texts
BnF fr. 185	133
BnF NAF 23686	98
BnF fr. 23117	98
BnF fr. 6447	92
Chantilly 734	87
Basel cl 102	85
BnF fr. 411	63
BnF fr. 412	61
BnF fr. 17229	61
BSG 588	47

(a) Number of texts per manuscript.



(b) Distribution of texts size, as number of token, per manuscript.

Figure 2: Description of the generated output corpora

in case of manuscripts with very few lives[17], (2) manuscripts with injected candidates from other ones (BnF fr. 412, Chantilly 734 and BnF fr. 185).

Evaluation loop

To determine whether Saint Lambert can be attributed to Wauchier de Denain, we apply the BDI's approach to each manuscript under three preprocessing schemes – raw, tokenized, and lemmatized – provided Wauchier has at least four texts in that codex. In every case, we extract up to two feature sets: (1) full tokens (words for raw and re-tokenized, lemmas for lemmatized) and (2) 3-gram affixes (only for raw and re-tokenized texts).

For a given combination of preprocessing and feature set, we remove one of Wauchier's texts as the test text, denoted T_W, and group the remaining Wauchier texts into the candidate set C_W. We then randomly split all non-Wauchier texts into two equal halves: impostors $I_{\neg W}$, which challenge the attribution, and negatives $T_{\neg W}$, which serve as precision benchmarks.

Next, every text is segmented into overlapping chunks of length $L = 2500$ tokens with a sliding

[17] BSG 588 has no lives of Wauchier outside of the hypothetical *Vie de Saint Lambert*; BnF, fr. 6447 has only three; BnF Fr. 17229 has only one.

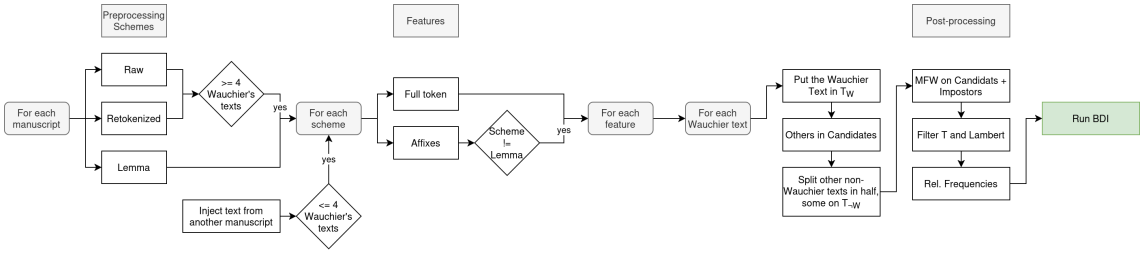


Figure 3: BDI Experiment pipeline, once pre-filtering has been applied.

window of $O = 1250$ tokens. We chose L based on the median text lengths in our collection, ensuring robust sample sizes. From the union $C_W \cup I_{\neg W}$, we select the top S most frequent features (with $S \in 200, 250, 300$). We then reduce T and Lambert to these features, and compute relative frequencies per text.

Finally, we run the BDI algorithm as in Nagy's study–1000 bootstrap iterations, Ružička's distance metric, and a 35% feature retention. For each iteration, we use all of T as well as Lambert as queries against C_W, recording the maximum statistical likelihood (SL) for *Vie de Saint Lambert*, comparing it to the SL distribution of W_{test}, as well as each run's distance. Repeating this procedure across all held-out texts and corpus variants yields both an accuracy measure of Wauchier's internal consistency and a relative likelihood for attributing Saint Lambert to him.

The Lambert's situation again ?

As *Saint Lambert* represents a particularly fortuitous discovery, other serendipitous finds cannot be ruled out.[18] To screen for these, we first process the entire corpus in its lemmatized form, supplementing any subset with fewer than four Wauchier texts by injecting the lemma corresponding to BnF, fr. 412. We then run BDI on each anonymous manuscript against both the complete set of Wauchier candidates (C_W) and the non-Wauchier impostors ($I_{\neg W}$). If an anonymous text achieves a statistical likelihood (SL) above 80%, we provisionally mark it as a potential Wauchier match and remove it from further saint Lambert attribution tests.

The pre-filtering step identified at least seven anonymous lives that matched Wauchier's profile in two or more manuscripts – three of which showed strong consistency across more than four manuscripts. In addition, 26 texts were excluded based on a single-manuscript match. Several of these texts appear to be unique to their respective manuscripts, suggesting that such isolated matches may result from manuscript-specific stylistic artefacts rather than true authorial overlap.

4 Results and interpretation

4.1 Statistical likelihood

Attribution of W_T remains at 100% statistical likelihood, meaning BDI works on Wauchier. *Saint Lambert*'s results (Table 5) tend to support its attribution to Wauchier de Denain, with generally high scores – often close to 100 percent – on 8 out of 10 manuscripts. While retokenization increases scores for words and affixes (except in one case), lemmas generally yield lower scores, possibly due to noise introduced by automatic lemmatization (which was trained only on synthetic data). High scores are also found in two of the three manuscripts with a limited number of texts by Wauchier de Denain, namely BSG 588 and BnF fr. 6447, whose *Lambert* shows greater affinity with Chantilly's (E) texts rather than with those in BnF fr. 412 (C) or BnF fr. 185 (G).

Two manuscripts mark an exception: BnF fr. 17229 (D) and BnF fr. 23117 (F). Previous research on the *Li Seint Confessor* [46; 49] has shown that BnF fr. 17229 includes a version of the

[18] Early experimental trials even identified several anonymous texts whose profiles closely resembled those of Wauchier's works.

Life of Saint Martin from the main branch of the textual tradition, but it contains no other works attributed to Wauchier. It is therefore plausible that the integration of Lives from different sources, along with the influence of scribal practices, may have introduced noise into the stylometric signal and blurred the attribution, yet BSG 588 (B) and BnF fr. 6447 (D as well) do not show this variation.

Regarding the second manuscript, BnF fr. 23117, previous analysis of *Saint Martin* has shown evidence of contamination between the two main branches of the textual tradition [49]. Most importantly, there is a shift in hands at folio 228, meaning that *La Vie de saint Lambert* was not copied in the manuscript by the same hand as the other Lives from the *Seint Confessor*. However, a quick check of the ATR output showed no significant variation in text quality or scribal features.[19] Both hybridizations may have disrupted the authorial signal and complicated the stylometric attribution. However, manuscripts G and L, which also exhibit hybridization in the composition of their Lives but not in their copying, do not display this phenomenon. Further investigation, including a full collation of the *Vie de saint Lambert* across the relevant witnesses, will be necessary to clarify the extent and nature of this interference.

Meyer's Family		B	C		D	D1	E		F	G	L
Manuscript		BSG 588	Fr. 412	Fr. 411	Fr. 17229	Fr. 6447	Chantilly, 734	Basel, cl 102	Fr. 23117	Fr. 185	NAF 23686
Raw	affixes		100	100			92	99	0	96	100
	words		100	96			97	94	0	94	99
Retokenized	affixes		100	100			100	100	0	98	100
	words		100	99			98	96	0	98	98
Lemmatized	words		98	99			97	87	26	95	93
+ Fr. 412	words	96			61	92					
+ Fr. 185	words	92			61	92					
+ Chantilly	words	97			40	98					

Table 5: Maximal statistical likelihood estimates attributing Lambert to Wauchier, based on Nagy's BDI function. Manuscripts lacking a sufficient number of saints' lives are evaluated solely through their lemma, with data from BnF Fr. 412 integrated to compensate for missing content. The scores shown reflect the highest possible classification as Wauchier (always 100%) for the reference text. All "fr." or "NAF" manuscripts are from the BnF. 300 MFW

4.2 Kernel Density Estimation's interpretation

KDE Peak on x	Implications
$x > 0$	Text is most likely authored by the candidate.
$x = 0$	Text is equally distant from both the candidate's texts and the impostors, implying that the text does not share an author with texts present in the corpus
$x < 0$	Text is more similar to one of the impostors than to the candidate, and is therefore most likely authored by the impostor that attracted it.

Table 6: Interpretation framework of Nagy's BDI KDE plots.

The KDE plots are much less definitive than the statistical likelihood, according to the reading framework of Nagy's (see Table 6).[20] While we see that T_W are correctly attracted by their pairs, we see *Vie de Saint Lambert* only partially overlapping the wave of T_W with small portions of its

[19] CER and abbreviation rates vary by approximately 1%; slight differences in the use of n, m, or r likely stem from ATR misreadings of downstrokes.

[20] The number of plots generated by the study is important (more than 30 generated plot, for which up to 9 KDE plot are produced), and we are not able to represent them all in the paper. We present a selected few, prototypes of the different situation we can find. All plots are however available in the pipeline repository, under `/data/plots`.

left roots starting before or around $x = 0$. Of the two manuscripts found to have negative statistical likelihood, 23177 is definitely on the negative side of the plot, meaning that the author of *Saint Lambert* in fr. 23117 would be attracted by another text in the corpus, which should not be of Wauchier because of our pre-filtering.

Yet, it seems rather improbable that *Saint Lambert* would not be drawn into the negative portion of the kernel. An attraction to the left, *i.e.*, in the negative distance range, would indicate that one of the anonymous impostors is the author of the text, or, as Nagy noted in the review of our paper, that a combination of texts within the impostor set may collectively attract the target text through overlapping feature proximities. When we examine the majority of tested impostors ($T_{\neg W}$), they tend to cluster on the left, suggesting that most texts appear in thematic or stylistic pairs within the legendiers. Both versions of E show two peaks in the kernel density plot in the positive range, indicating that some Lives are more "Wauchierish" than others from BDI's "point of view".

We propose three hypotheses, ranked from the most to the least probable based on the observed results:

1. *Saint Lambert* is less strongly attracted to Wauchier's texts because it is thematically more distant, and most importantly, because it does not belong to either the *saints Peres d'Egipte* or *Confessors* collections.

2. *Saint Lambert* shows a weaker stylistic affinity with Wauchier's texts, possibly because it is an earlier work or at least one composed for a different audience and in a different context, as suggested for instance by the absence of verse passages.

3. *Saint Lambert* is not authored by Wauchier but remains close to his style, potentially because it was written using the same sources, in the same geographic and cultural environment, by an author whose style closely resembles that of Wauchier.

Conclusion

The results mostly confirm the initial "discovery" in [7]. While our stylometric analyses strengthen the attribution of *Saint Lambert* to Wauchier de Denain, we remain cautious: BDI, though effective in clustering and detecting stylistic affinities, does not offer the highest precision for individual attribution decisions, and a margin of error persists. Notably, a few genuine texts by Wauchier in the test set were occasionally misclassified as non-Wauchier, underscoring the method's limitations—particularly in the face of noise and sparse data.

Nonetheless, when these stylometric findings are combined with manuscript and literary evidence, a coherent picture begins to emerge. Several converging signals suggest that *Saint Lambert* may indeed be written by Wauchier, or at least the work of a closely affiliated author operating within the same geographic, cultural, and intellectual milieu. The stylistic proximity is too consistent to dismiss as coincidence, even if it cannot yet be certified with full certainty.

Going forward, the next step is clear: to collate and publish a critical edition of the text of *Vie de Saint Lambert*, allowing for more in-depth philological study and comparative textual analysis. Special attention should also be paid to those sections and texts excluded during preprocessing, which may contain valuable clues. If nothing else, this study highlights the inherent difficulty of applying authorship verification techniques to medieval corpora, where anonymous transmission is the norm rather than the exception, and where authorial identity is often blurred by layers of textual transmission and scribal intervention.

While our stylometric resources are, for now, exhausted, the evidence we have gathered leans, cautiously but positively, toward affirming Wauchier's authorship of *Saint Lambert*.

References

[1] Aguilar, Sergio Torres. “TRIDIS: A Comprehensive Medieval and Early Modern Corpus for HTR and NER”. In: *arXiv preprint arXiv:2503.22714* (2025).

[2] Beullens, Pieter, Haverals, Wouter, and Nagy, Ben. “The Elementary Particles. A Computational Stylometric Inquiry into the Medieval Greek-Latin Aristotle”. In: *Mediterranea. International Journal on the Transfer of Knowledge 9* (2024), pp. 385–408. DOI: 10.21071/mijtk.v9i.16723.

[3] Boillet, Mélodie, Kermorvant, Christopher, and Paquet, Thierry. “Robust text line detection in historical documents: learning and evaluation methods”. In: *International Journal on Document Analysis and Recognition (IJDAR) 25*, no. 2 (2022), pp. 95–114. DOI: 10.1007/s10032-022-00395-7.

[4] Burrows, John. “‘Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship”. In: *Literary and Linguistic Computing 17*, no. 3 (Sept. 2002), pp. 267–287. DOI: 10.1093/llc/17.3.267.

[5] Cafiero, Florian and Camps, Jean-Baptiste. “Why Molière most likely did write his plays”. In: *Science advances 5*, no. 11 (2019), eaax5489. DOI: 10.1126/sciadv.aax548.

[6] Camps, Jean-Baptiste, Clérice, Thibault, Duval, Frédéric, Ing, Lucence, Kanaoka, Naomi, and Pinche, Ariane. “Corpus and Models for Lemmatisation and POS-tagging of Old French”. 2021. URL: https://arxiv.org/abs/2109.11442.

[7] Camps, Jean-Baptiste, Clérice, Thibault, and Pinche, Ariane. “Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer’s hagiographic hypothesis”. In: *Digital Scholarship in the Humanities 36*, no. Supplement_2 (Oct. 2021), pp. ii49–ii71. DOI: 10.1093/llc/fqab033.

[8] Camps, Jean-Baptiste, Salvati, Benedetta, Freijedo, Gonzalo, Bian, Donghan, Drouet, Gaëtan, Gaglione, Eglantine, Guidi, Émilie, Macedo, Carolina, Zribi, Yaelle, and Cafiero, Florian. “The Authorship of the Works of Chrétien de Troyes: a Stylometric Examination”. In: *DH Benelux 2024*. 2024.

[9] Clérice, Thibault. “Building a cross-lingual dataset from medieval manuscript text recognition Challenges and outcomes of CATMuS.” London, United Kingdom: Computational Humanities Research Group Spring Seminar, 2025. URL: https://inria.hal.science/hal-05150349.

[10] Clérice, Thibault. “Evaluating deep learning methods for word segmentation of Scripta Continua texts in old French and Latin”. In: *Journal of Data Mining and Digital Humanities 2020* (2020).

[11] Clérice, Thibault and Glaise, Anthony. “Twenty-One* Pseudo-Chrysostoms and more: authorship verification in the patristic world”. In: *Computational Humanities Research Conference 2023*. Vol. 3558. Paris, France, 2023, pp. 222–247. URL: https://ceur-ws.org/Vol-3558/paper2277.pdf.

[12] Clérice, Thibault, Pinche, Ariane, Vlachou-Efstathiou, Malamatenia, Chagué, Alix, Camps, Jean-Baptiste, Levenson, Matthias Gille, Brisville-Fertin, Olivier, Boschetti, Federico, Fischer, Franz, Gervers, Michael, et al. “CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond”. In: *International Conference on Document Analysis and Recognition*. Springer. 2024, pp. 174–194. DOI: 10.1007/978-3-031-70543-4_11.

[13] Delarue, Dominic E. *Die Legendare aus der Rue Neuve Nostre Dame: Dispositio und Bildformeln in der Pariser Buchmalerei, 1325–1348*. Vol. 23 / 16. Corpus of Illuminated Manuscripts. Low Countries Series. 2021.

[14] Delisle, Léopold. *Le cabinet des manuscrits de la Bibliothèque impériale: étude sur la formation de ce dépôt comprenant les éléments d'une histoire de la calligraphie de la miniature, de la reliure, et du commerce des livres à Paris avant l'invention de l'imprimerie*. Impr. impériale nationale, 1868.

[15] "Vie de Saint Lambert". In: *L. Grandmont-Donders, imprimeur-libraire* (1890), ed. by Joseph Demarteau, p. 70. URL: https://donum.uliege.be/handle/2268.1/12667.

[16] Denain, Wauchier de. *La vie de Mon Signeur Seint Nicholas le Beneoit confessor*, ed. by John Jay Thompson. Droz, 1999.

[17] Eder, Maciej. "Boosting word frequencies in authorship attribution". In: *Computational Humanities Research Conference 2023*. Vol. 3290. Antwerp, Belgium, 2022, pp. 387–397. URL: https://ceur-ws.org/Vol-3290/long_paper5362.pdf.

[18] Eder, Maciej. "Rolling stylometry". In: *Digital Scholarship in the Humanities 31*, no. 3 (Apr. 2015), pp. 457–469. DOI: 10.1093/llc/fqv010.

[19] Eder, Maciej. "Taking stylometry to the limits: Benchmark study on 5,281 texts from Patrologia Latina". In: *Digital humanities 2015: conference abstracts*. 2015, pp. 1919–1924.

[20] Gorman, Robert. "Author identification of short texts using dependency treebanks without vocabulary". In: *Digital Scholarship in the Humanities 35*, no. 4 (Oct. 2019), pp. 812–825. DOI: 10.1093/llc/fqz070.

[21] Guillot, Céline, Heiden, Serge, and Lavrentiev, Alexei. "Base de français médiéval: une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique". In: *Diachroniques. Revue de Linguistique française diachronique* , no. 7 (2018), pp. 168–184.

[22] Hernández-Lorenzo, Laura and Tello, José Calvo. "Stylometric Evaluation of Parameters and Distance Measures for Hispanic Texts". In: *Digital Humanities in Medieval and Early Modern Spanish Texts*. Routledge, 2025, pp. 163–188. DOI: 10.4324/9781003393771.

[23] Hochuli Dubois, Paule. "Notice de Genève, Bibliothèque de Genève, Comites Latentes 102". In: *Bibliothèque de Genève, pour e-codices* (2016).

[24] Kestemont, Mike, Stover, Justin, Koppel, Moshe, Karsdorp, Folgert, and Daelemans, Walter. "Authenticating the writings of Julius Caesar". In: *Expert Systems with Applications 63* (2016), pp. 86–96. DOI: 10.1016/j.eswa.2016.06.029.

[25] Khanam, Rahima and Hussain, Muhammad. "YOLOv11: An Overview of the Key Architectural Enhancements". 2024. arXiv: 2410.17725 [cs.CV]. URL: https://arxiv.org/abs/2410.17725.

[26] Kiessling, Benjamin. "Large Multilingual ATR Models and Humanities Practice". In: *Workshop SCOOP - Source Codes of the Past*. Princeton , NJ, United States, June 2025. URL: https://inria.hal.science/hal-05150070.

[27] Kiessling, Benjamin. "Version 5 of the Kraken ATR Engine for the Humanities". In: *ICDAR 2025*. Wuhan, China, 2025, pp. 443–458. DOI: 10.1007/978-3-032-04624-6_26.

[28] Koppel, Moshe and Winter, Yaron. "Determining if two documents are written by the same author". In: *Journal of the Association for Information Science and Technology 65*, no. 1 (2014), pp. 178–187. DOI: 10.1002/asi.22954.

[29] Kupper, Jean-Louis and George, Philippe. *Saint Lambert: de l'histoire à la légende*. 2006.

[30] Kurth, Godefroid (1847-1916). *Etude critique sur Saint Lambert et son premier biographe*. July 2023. URL: https://donum.uliege.be/handle/2268.1/12670.

[31] Labie-Leurquin, Anne-Françoise. "Composition, usage et diffusion du légendier picard". In: *Des saints et des livres : Christianisme flamboyant et manuscrits hagiographiques du Nord à la fin du Moyen Âge (xi-xiiie siècles)*. Hagiologia 17. 2021, pp. 79–133.

[32] Manjavacas, Enrique, Clérice, Thibault, Nédey, Oriane, Kestemont, Mike, and Akos, Kadar. "(Pa)PIE: A Framework for Joint Learning of Sequence Labeling Tasks". Version 0.5.0. Apr. 2023. URL: https://github.com/lascivaroma/PaPie.

[33] Mattingly, William. "biglam/medieval-manuscript-yolov11 · Hugging Face". July 22, 2024. URL: https://huggingface.co/biglam/medieval-manuscript-yolov11.

[34] Meyer, Paul. "Légendes hagiographiques en français". In: *Histoire littéraire de la France*. Vol. 33. Paris: Paris : Imprimerie nationale, 1906, pp. 328–458. URL: http://archive.org/details/histoirelittra33riveuoft.

[35] Nagy, Ben. "Bootstrap Distance Imposters: High precision authorship verification with improved interpretability". In: *Proceedings of the Computational Humanities Research Conference 2024*. Computational Humanities Research 2024, ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. Vol. 3834. CEUR Workshop Proceedings. Aarhus, Denmark: CEUR, Dec. 4, 2024, pp. 482–493. URL: https://ceur-ws.org/Vol-3834/#paper61.

[36] Nagy, Benjamin. "Metre as a stylometric feature in Latin hexameter poetry". In: *Digital Scholarship in the Humanities 36*, no. 4 (Feb. 2021), pp. 999–1012. DOI: 10.1093/llc/fqaa043.

[37] Philippart, Guy. *Les Légendiers latins et autres manuscrits hagiographiques*. Brépols, 1977.

[38] Pinche, Ariane. "Generic HTR Models for Medieval Manuscripts. The CREMMALab Project". In: *Journal of Data Mining & Digital Humanities Historical Documents and automatic text recognition* (Oct. 2023). DOI: 10.46298/jdmdh.10252.

[39] Pinche, Ariane et al. "CATMuS Medieval". Version 1.6.0. Mar. 2025. DOI: 10.5281/zenodo.15030337.

[40] Pinche, Ariane et al. "CATMuS Medieval". Version Number: 1.6.0. Mar. 2025. DOI: 10.5281/zenodo.15030337.

[41] Rebora, Simone, Herrmann, J Berenike, Lauer, Gerhard, and Salgaro, Massimo. "Robert Musil, a war journal, and stylometry: Tackling the issue of short texts in authorship attribution". In: *Digital Scholarship in the Humanities 34*, no. 3 (Oct. 2018), pp. 582–605. DOI: 10.1093/llc/fqy055.

[42] Rouse, Richard H. and Rouse, Mary A. *Manuscripts and their makers: Commercial book producers in medieval Paris, 1200-1500*. 2000.

[43] Simeray, Françoise. *Le scriptorium et la bibliothèque de l'abbaye Saint-Amand*. Thèse diplôme d'archiviste-paléographe. 1990.

[44] Stones, Alison. *Gothic Manuscripts 1260–1320*. Vol. 1-2. A Survey of Manuscripts Illuminated in France. 2013.

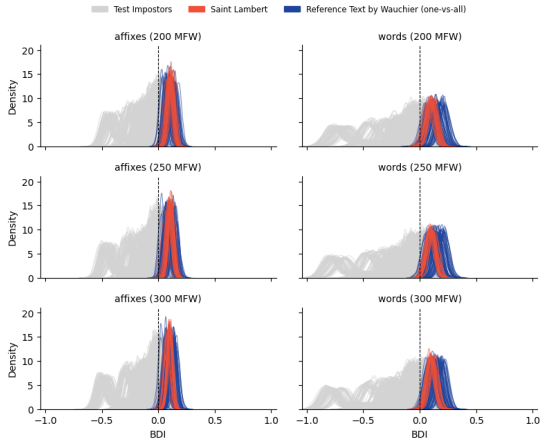
[45] Tarride, Solène, Schneider, Yoann, Generali-Lince, Marie, Boillet, Mélodie, Abadie, Bastien, and Kermorvant, Christopher. "Improving automatic text recognition with language models in the pylaia open-source library". In: *Document Analysis and Recognition - ICDAR 2024*. Cham, 2024, pp. 387–404. DOI: 10.1007/978-3-031-70549-6_23.

[46] Thompson, John Jay. *From the translator's worktable to the predictor's lectern: The work of a thirteenth-century author, Wauchier de Denain*. English. PhD in Candidacy for the Degree of Doctor of Philosophy. Yale: Yale University, 1993. (Visited on 05/21/2018).

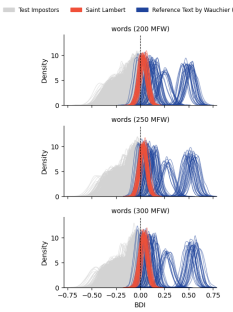
[47] Vandyck, Caroline and Kestemont, Mike. "Abbreviation application: a stylochronometric study of abbreviations in the oeuvre of Herne's Speculum Scribe". In: *Proceedings of the Computational Humanities Research Conference 2024*. Vol. 3834. 2024, pp. 881–891. URL: `https://ceur-ws.org/Vol-3834/paper15.pdf`.

[48] Wauchier de Denain. *L'histoire des moines d'Égypte*, ed. by Michelle Szkilnik. Genève: Droz, 1993.

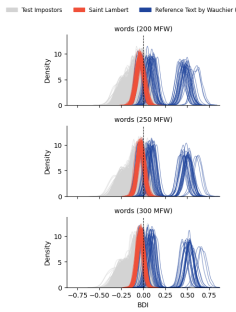
[49] Wauchier de Denain. *Li seint confessor*, ed. by Ariane Pinche. Paris, France: Honoré Champion éditeur, 2024.



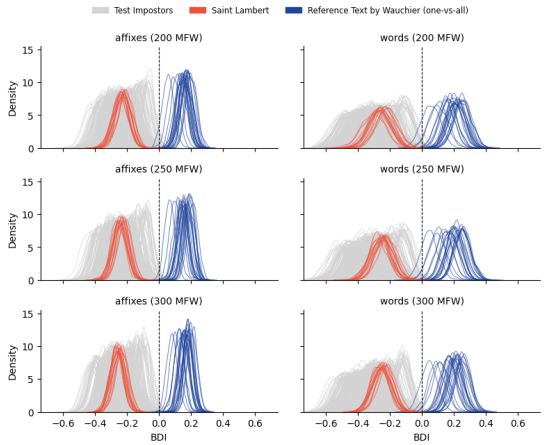
(a) BnF fr. 412, Preprocessing scheme = Retokenized



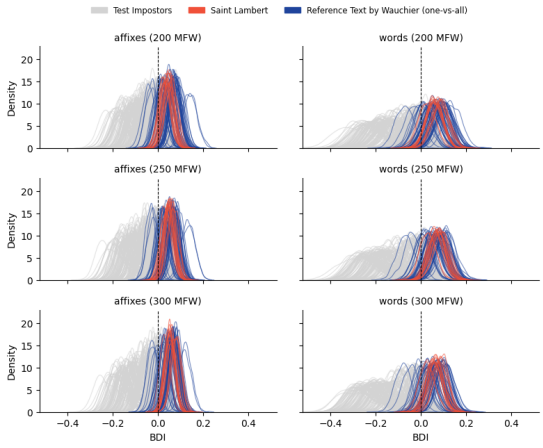
(b) BnF, fr. 6447 with lives injected from Chantilly 734, Preprocessing scheme = lemma



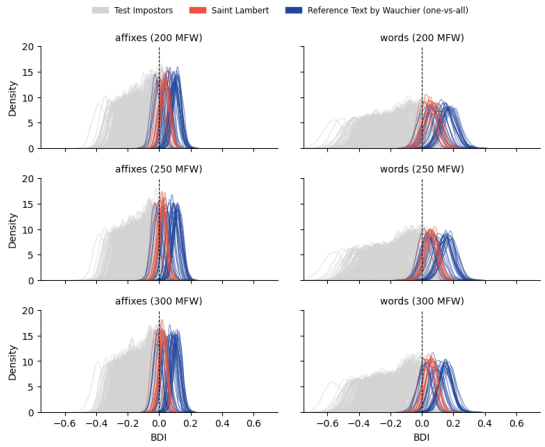
(c) BnF fr. 17229 with lives injected from Chantilly 734, Preprocessing scheme = lemma



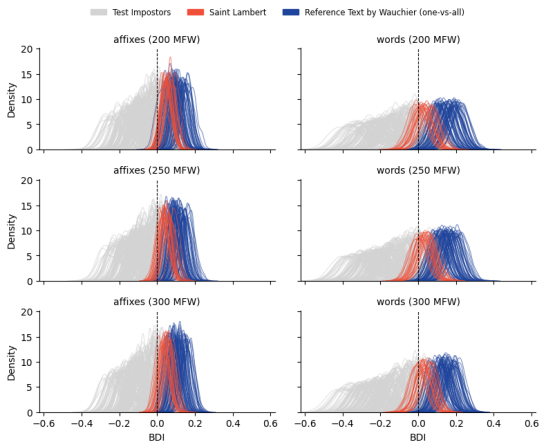
(d) BnF, fr. 23117, Preprocessing scheme = Retokenized



(e) BnF, fr. 23686, Preprocessing scheme = Retokenized



(f) Chantilly 734, Preprocessing Scheme = Raw



(g) BnF fr. 411, Preprocessing Scheme = Raw

Figure 4: Distribution plots for various manuscripts and preprocessing schemes. (c) and (d) are negatives, (b) and (e) show some relatively good positive KDEs. Blue = T_W, gray = $T_{\neg W}$ and orange = Lambert.