CoMMA, a Large-scale Corpus of Multilingual Medieval Archives

Thibault Clérice[1], **Simon Gabay**[2], **Malamatenia Vlachou-Efsthatiou**[3], **Ariane Pinche**[4], **Benoît Sagot**[1]
(1) Inria, Paris, France. name.surname@inria.fr
(2) Université de Genève, Geneva, Switzerland. name.surname@unige.ch
(3) IRHT/CNRS – ENPC, Paris, France. name.surname@enpc.fr
(4) CNRS, Lyon, France. name.surname@cnrs.fr

Abstract

We present CoMMA, a large-scale corpus of medieval manuscripts produced through automatic text recognition. The corpus contains around 3.3b tokens drawn from more than 32,700 digitized manuscripts in Latin and Old French, harvested via IIIF. Unlike other resources, it is made of raw, non-normalized text enriched with layout analysis in various formats. We describe the pipeline used for large-scale acquisition and processing, and report quantitative and qualitative evaluations (average CER 9.7%). The resulting resource supports multiple use cases, from pretraining language models to corpus linguistic on historical languages and digital humanities applications.

Keywords: Automatic Text Recognition, Medieval manuscripts, Latin, French, Digital humanities, Corpus

1. Introduction

Estimates suggest that around 1.3 million manuscripts from the Latin West produced before 1800 still survive today (Buringh, 2010). The Biblissima IIIF-Collection (Morlock et al., 2025) currently provides metadata for approximately 143,000 digitized manuscripts, including some 65,000 described as Latin and 18,500 in premodern varieties of French. Despite this wealth of material, accessible large-scale datasets for Old French and medieval Latin remain scarce. For computational linguists and NLP practitioners, most available resources are limited in scope, editorialized, and heavily normalized for readability (including but not limited to punctuation, capitalization, word segmentation, and finally abbreviation resolution), which obscures the original manuscript practices of abbreviation, "misspellings", and orthographic variability.

For Old and Middle French, spelling variation is not accidental but structural. Prior to the orthographic stabilization that spread gradually from the Renaissance (Gabay and Clérice, 2024), spelling practices themselves functioned as markers of place and period of production. Dees (1985) pioneering quantitative work on Old French dialects highlighted these regional patterns. Yet, Morin (2006) has shown that reliance on critical editions introduced distortions: in "Floovant"—one of the texts produced by Dees and later captured in the Nouveau Corpus d'Amsterdam (NCA Stein and Kunstmann 2006)—a corpus resulting of Dees work, Morin showed that the ‹ȝ› sign from the manuscript is expanded as *et* ('and') in the edition while used in the coordinating function, ignoring the fact that ‹ȝ› is also used in place of *e* or *a* in prepositional contexts (modern French *à* 'at'), leading him to believe that the normalization *e* is as valid as *et*, potentially changing relative frequencies of both forms. Printed editions and their digitization, which comprise the vast majority of the digital corpora available online, conceal at best – erase at worst – both the abbreviation practice and the choice made by the editor, which sometimes reinforce or suppress local spellings. For computational philology and corpus linguistics (CL), this raises a key issue: statistical conclusions drawn from such corpora may be skewed by this practice.

Recent advances in automatic text recognition (ATR) have made it possible to address this challenge at scale. For medieval manuscripts in Western languages, recognition accuracy is increasingly limited, not by models but by the availability of ground-truth data, with current error rates often below 10%. Initiatives such as the CATMuS Medieval dataset (Clérice et al., 2024) have provided training material across 9 centuries (8-16th) and ten western languages written in Latin script, covering a range spanning from administrative registers to hagiographic texts. This dataset opens the way to large-scale corpus creation directly from manuscript images and keeps abbreviations in its transcription ground truth.

In this paper, we present the CoMMA corpus, a new large-scale resource built from 32,763 digitized manuscripts, totaling 3.33 billion tokens. To our knowledge, this is more than 12 times the size of the largest open Latin corpus currently available (2.7B vs. 226M, *Corpus Corporum*, Roelli and Ctibor, 2024), and represents more than 40 times the open Old French as well (516.8M vs 11.2M). Unlike most existing corpora, the CoMMA corpus retains abbreviation signs, preserving the diversity of manuscript practices.

Our main contributions are threefold:

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7034–7045
11–16 May 2026.

- **A scalable pipeline** for corpus construction from manuscript digitizations, combining ATR with layout analysis using the SegmOnto controlled vocabulary (Gabay et al., 2024);
- **The CoMMA corpus itself**, comprising both raw text and semi-structured output using XML-TEI, openly available for NLP and corpus linguistics research;
- **Evaluation and applications**, including character error rate (CER) benchmarks, lexical overlap measures, and example use cases in language modelling, historical linguistics, and manuscript layout analysis.[1]

The remainder of this paper is structured as follows. In section 2, we review related work on historical language corpora and large-scale ATR initiatives. In Section 3, we describe the corpus sources, coverage, and processing pipeline. In section 4, we provide metrics regarding the corpus and evaluate its quality and metadata. In section 5, we present three example applications on the corpus: (1) analysis of the evolution of book layout practices, (2) analysis of the evolution of scribal practices, and (3) pretraining of a ModernBERT model (Warner et al., 2024) for semantic manuscript exploration. Section 6 concludes with future directions for extending and annotating the corpus.

2. Related Work

2.1. Large Corpora in Old French and Latin

Corpus building for Latin traces back to the pioneering work of Roberto Busa, one of the first to systematically digitize textual resources (Busa, 1980). Since then, numerous large-scale textual corpora have been produced, ranging from linguistically annotated datasets, such as LASLA's dataset—the largest annotated Latin corpus—to simply readable text collections (Denooz, 1996). Many of the largest corpora, however, remain behind paywalls (e.g., LLT-A and LLT-B, Centre Traditio Litterarum Occidentalium 2025, FranText, Montémont 2020) or accessible only through controlled reading environments (e.g., NCA at ATILF, Stein and Kunstmann 2006). This restricted access limits its potential uses, from vectorization to pretraining language models. Other open projects, such as CEMA Perreaux (2021), while freely accessible, do not yet support large-scale computational usage and often require manual download of individual files, making large-scale processing cumbersome.

Openly licensed and easy-to-download corpora for Latin and Old French do exist. Examples for Latin include the Perseus Digital Library for the classical period (Crane et al., 2021b), DigilibLT for late antiquity Latin (Tabacco et al., 2021), and the CSEL through the Open Philology Project (Crane et al., 2021a) for Christian Latin. The Corpus Corporum (Roelli and Ctibor, 2024) provides a unified entry point to many of these resources. For Old French, the research community is smaller, but several open-access corpora exist, including the Base de Français Médiéval (BFM, Guillot et al. 2018) as well as DocLing for charters (Glessgen, 2017). These resources amount to only a few hundred million tokens: 10 million tokens for Old French, 220 million for Latin.

The development process of the first Latin BERT model (Bamman and Burns, 2020) illustrates the limitations of existing corpora: to reach sufficient scale for pretraining, the authors relied on uncorrected OCR from `Archive.org`, adding 560 million tokens from uncurated and uncorrected OCR to a smaller yet curated corpus of 81M tokens produced by various Digital Humanities (DH) projects.

Moreover, a common characteristic of these projects is that they mostly rely on printed editions as their primary source. While convenient, this practice smooths the inherent diversity present in manuscript transmission and often normalizes original writing practices. For instance, in the document identified as RM 1291 06 09 01 from DocLing—a project with digital born editions—the manuscript spelling ‹cest asauoir q̣lsũt› is normalized as "C'est à-savoir q*u*'i*l* su*n*t" (italic identifies resolved abbreviation). This version of the text obscures the manuscript's variability: there is no indication that the normalization of spacing (or retokenization) originates with the scientific editor and thus constitutes an editorial intervention. Such interventions underscore the challenge of capturing genuine manuscript variation — and, by extension, linguistic variation — in computational corpora.

Recent advances in ATR have enabled the creation of large-scale corpora from historical manuscripts. Notably, the Open Islamicate Texts Initiative (OpenITI, Miller et al. (2018)) provides a machine-actionable corpus of Persian and Arabic texts, while projects such as MiDRASH (Stoekl Ben Ezra et al., 2025) aim to release similar corpora for other languages.

2.2. Layout Analysis

Layout analysis is the process of detecting and describing the structural elements of a page, such

[1] Pipeline: `https://github.com/DEFI-COLaF/CoMMA-pipeline`. Models and data: `https://huggingface.co/comma-project`. Reading environment: `https://comma.inria.fr`.

as text regions and lines, to support automatic text recognition. While this task is well studied for modern documents, medieval manuscripts pose additional challenges due to irregular page structures, variable writing supports, decorated initials, and complex marginalia (Clausner et al., 2017; Grüning et al., 2018). Recent state-of-the-art approaches rely on deep learning, with convolutional and region-based models (e.g., Faster R-CNN, Mask R-CNN) as well as fully convolutional networks for line detection and region classification (Yang et al., 2017; Oliveira et al., 2018). YOLO-based models have demonstrated efficiency for both region and line detection (Zhao et al., 2024).

The SegmOnto controlled vocabulary (Gabay et al., 2024) provides a standardized framework to label main text blocks, marginal notes, intralinear glosses, and other frequent codicological features. Using SegmOnto classes allows for extracting textual resources from the main body of text while ignoring marginal or noise-inducing lines surrounding it, such as running titles. CATMuS (Clérice et al., 2024) adopted this vocabulary to ensure consistent annotation across heterogeneous manuscript sources and provided a corresponding segmentation dataset.

2.3. Extracting Text from Medieval Manuscripts

ATR has progressed rapidly from Convolutional Recurrent Neural Networks (CRNNs) in tools such as Kraken (Kiessling, 2022) and Pylaia(Tarride et al., 2024) to transformer-based approaches (Diaz et al., 2021; Li et al., 2023) and multimodal LLMs (Crosilla et al., 2025). Yet most work on handwritten documents focuses on 19th–20th-century handwriting in standardized languages, whereas medieval sources exhibit highly variable spelling — both through the lack of consistent orthographic rules and through extensive abbreviation practices. Transformer-based models, large language models (LLMs), and even some CRNN-based tools, such as PyLaia, which can integrate KenLM language models (Heafield, 2011), rely on a language model to interpret visual cues. This modelling helps disambiguate between possible transcriptions when the written trace is difficult to decipher, and as a consequence can overnormalize content.

Ground truth creation follows two traditions. *Pre-editorialized transcription* normalizes abbreviations, spelling, and capitalization, treating automatic transcription as the first stage of editing (Aguilar, 2025). By contrast, *graphematic transcription* preserves abbreviation systems while collapsing only graphic variants (also called *allographs*, e.g., ‹ſ›/‹s›), as codified by Pinche (2022). The CATMuS corpus exemplifies the latter, offering over 200,000 lines from 300 manuscripts across 10 languages and 9 centuries.

However, both strategies face persistent risks. Graphematic training produces expected character-level confusions (e.g., ‹ri›/‹n›), while pre-editorialized models replicate the over-normalization and hallucination effects documented by Torterolo-Orta et al. (2025) and Bottaioli et al. (2024), including silent abbreviation resolution and catastrophic semantic substitutions (e.g., ‹dia℈› transcribed as ‹diaconus› ‘deacon’ instead of ‹diabolus› ‘devil’, see Aguilar 2024). Thus, despite advances with transformers and Multi-modal Large Language Models, normalization and normalization-induced hallucination remain critical challenges, which have led us to prefer the CATMuS approach for building datasets where errors are more readily detectable.

3. Producing the Corpus

3.1. Source and Coverage

The CoMMA corpus is drawn from a query on the Biblissima portal.[2] The query targeted manuscripts in Latin or French, produced between 800 and 1550, and available from five digitization providers with permissive terms of use and sufficient download throughput: ARCA, Gallica, E-Codices, the Bodleian Library and the Bayerische Staatsbibliothek. This query returned more than 49,443 candidate manuscripts.

Several types of metadata-induced noise commonly arise in such large-scale harvesting, namely: duplication of manuscripts (through the duplication of digitization efforts between research projects such as ARCA and libraries), incorrect dating or language metadata (for example, Old Occitan and Old French are frequently both marked as “French” in BnF records.), unidentified composite manuscripts (fragments collated at a later date, often during 19th-century bindings, leading other metadata such as date to be inconsistent), and finally partial digitization (digitization in ARCA are sometime led only for codicological reasons and can focus for example on bindings or decorations).

The focus on these institutions introduces certain biases. Because our harvesting primarily targeted libraries, administrative charters are underrepresented compared to their actual abundance, as archives—where such documents are traditionally held in France and other countries—are less accessible. Unfortunately, the digitization and open release of these archival materials lag behind those of libraries. Another important bias comes

[2] Results were provided by the Biblissima+ team.

from the digitization strategies of the institutions, which may decide to prioritize certain documents over others in the context of thematic online exhibitions.

3.2. Production Pipeline

In order to produce the corpus from the 33,000 manuscripts, we developed a pipeline able to deal with both the scale and the instability of the downloading process. The workflow proceeds in three stages. First, the metadata manifest of each manuscript is retrieved and cached. Second, images are downloaded with a maximum height of 2,500 pixels, following the practice established in the CATMuS datasets, and are processed in parallel while respecting the constraints imposed by the providers. Then, layout analysis and recognition were carried out using two different libraries: YOLOv11 (Khanam and Hussain, 2024) for layout segmentation and Kraken for ATR, both outputting ALTO-XML. YOLOv11 was run with a model trained on CATMuS Medieval Segmentation open data, using the SegmOnto vocabulary for regions and lines (Mattingly, 2025). Kraken employed the CATMuS 1.6.0 model (Pinche et al., 2025). As a result of our model choices, our corpus follows the CATMuS guidelines, as closely as the models use apply them faithfully.

Quality safeguards were introduced at each stage, verifying: (1) image readability, (2) XML existence (indicating that layout analysis was at least performed), (3) XML parsability, and (4) presence of text predictions within the XML. Any failure triggered reprocessing. Once a manuscript reached 100% page coverage, the XML files were archived along with a custom manifest recording page order and original IIIF manifest URI.

ALTO files were subsequently converted into TEI-XML pages, focusing exclusively on textual content (i.e., without spatial coordinates or links to the original image). A mapping was defined to align SegmOnto categories with TEI structures: `MainZone` was rendered as `<ab>` blocks; `DefaultLine` as plain text lines; and `HeadingLine` as `<span class="rubricated">`, while page change is indicated through `<pb>`. Other content was mapped to specific elements, such as `<note>` for marginal or interlinear text, and `<fw>` for catchwords or running titles. This design enables flexible exports: for instance, the main body of a manuscript can be obtained by simply ignoring `note` and `fw` tags, while it is equally possible to export only the notes. To facilitate usage, each manuscript is released with the TEI files, the corresponding ALTO files, JSON files with layout information and text from the original documents, and a pre-extracted main text version, sparing users from the need to write XML-handling scripts.

4. Resulting Corpus

4.1. A Corpus with Abbreviations

An essential feature of the CATMuS guidelines is the normalization of letter variants to ensure consistency across sources and languages. Abbreviations are preserved in their original form, encoded in Unicode Normalization Form Decomposed (NFD), and supplemented with special characters registered through the Medieval Unicode Font Initiative (MUFI). A limited number of characters from the Unicode Private Use Area are also employed, such as ‹;› (U+F1AC), defined in the MUFI, which—despite its resemblance to a semicolon—is conventionally used for abbreviating *-ed* or *-ibus* at the end of words. Graphical alternations that lack linguistic contrast, such as *u*/*v* or *i*/*j*, are collapsed into a single form (*u*/*i*).

The CATMuS guidelines also normalize spacing around word boundaries, as manuscript spacing is often ambiguous or inconsistent. Punctuation is restricted to two sentence-level signs (‹.› and ‹:›), but several additional content or typographic markers are retained: hyphenation (‹-›), content shifts (‹¶›), note insertions (‹ ̂›), and token separation (‹/›). To illustrate these conventions, Table 1 presents 10 randomly sampled lines from the main body of text in five Latin and five French manuscripts.

4.2. Statistics

Corpus Overview Out of the 49,443 queried digitizations, the corpus contains 32,763 distinct manuscripts. Over the 32,763 manuscripts, 25,270 are labelled as Latin only, 5,611 as French, 1,120 as bilingual, 413 as Italian, 288 as mixing either Latin or French with another language and 59 as trilingual. The average number of digitized images per unique manuscript is 172 ($\sigma = 175$), the number of lines per manuscript is 15,123 ($\sigma = 21,148$), while the average number of white-space-delimited tokens is 101,649 ($\sigma = 143,801$). The distribution of manuscripts and word counts across centuries is shown in Table 2 and 3. The total number of tokens is 3.3B.

Language Classification of Manuscripts To systematically categorize our manuscript data by language, we designed a rule-based classification using series of window-level language scores, with each window representing 1% of the tokens. We trained a FastText language identification model (Joulin et al., 2016) on the CATMuS Medieval dataset at the line level, using its ten predefined language classes. The training and development

	Latin	French
Readable	Sanguis : in altari sumit̃ ipse manẽs. fabrica nr̃a p ħ quod scd̓o sonuit fci ẽ Sulatis qs dn̄e placare munerib⁹ & int̃ cedẽtib; sistuis. potestis q᷒ int⁹ istus iinebiras diffinlior ad xxuii E. Obitus thome de benoist qui dedit fabrice huius ecc̊e	et depute nos Baillifs et Seneschaux sur les ¶mõseigneur quãt tes faiz ie recorde comptes du faut delas tresori/ en son propre nom p᷒g̃ Souspirant. ex criant ex ne ueuxe point tascher Q ui p᷒e ȝmere des honeure
Too noisy	AiU Tonclisio ebri et siguario dap a pgem⁹ dei fficie op̄a. nõg̃ sno psmeno ol reg deaẽ OParencũ omor l̊i̊ .ir̊. c̊å.ĥ	Iour de nommbir lan nulmy ourr ey Rend R mouerubut e Sl le puueast Re fio uons priiat fayr mi pomimuattan du dee uauy dal l heuants Regaois auspant

Table 1: 10 lines with different level of readability and abbreviation rate

Century	9	10	11	12	13	14	15	16	Total
Fr	0	0	8	24	511	796	2410	1862	5611
Fr, La	5	7	7	55	120	179	634	113	1120
Fr, La, Ot	0	1	0	5	10	10	23	10	59
Fr, Ot	0	0	0	1	11	11	20	11	54
It	0	0	0	0	2	54	208	149	413
La	1089	699	1153	4234	5659	4922	6133	1381	25270
La, Ot	14	10	12	21	17	30	93	37	234

Table 2: Distribution of manuscripts across centuries according to original metadata. *Ot* stands for Others.

Century	9	10	11	12	13	14	15	16	Total
Fr.			0.61	0.80	78.37	106.07	238.18	92.82	516.85
Fr, La	0.73	0.62	0.22	3.01	8.85	14.01	20.99	3.17	51.60
Fr, La, Ot		0.10		0.85	1.23	0.72	0.75	0.32	3.97
Fr, Ot					0.46	0.62	1.20	0.46	2.74
It					0.24	4.84	15.62	7.26	27.96
La	73.91	41.47	82.98	371.52	820.71	776.37	483.59	60.51	2711.06
La, Ot	1.04	0.62	0.46	2.05	2.09	4.13	4.08	1.64	16.11

Table 3: Token counts per century (M tokens). *Ot* stands for Others.

sets were used to train the model, while we used the test set to evaluate its performance. The resulting classifier achieved strong F1 scores for French and Latin (89.4% and 95.5%). Overall, the model attained an accuracy of 88.6% across the entire corpus.

For each manuscript, we computed smoothed summary features over a sliding window, including the mean and standard deviation for French, Latin, and an *Other* category ($P_{Other} = 1 - (P_{French} + P_{Latin})$), as well as the mean difference between Latin and French capturing their balance. These metrics formed the basis for a hierarchical rule-based classification scheme: manuscripts with highly dominant and stable French or Latin lines were labelled *Truly French* or *Truly Latin* ($\mu_{lang} > 0.9$ and $\sigma_{lang} < 0.05$). Those with a clear majority of one language but some variability were classified as *Massively French*, *Massively Latin* or *Massively Other* (threshold 0.7 and $\mu_{lang} > \mu_{\text{other}}$).

Manuscripts containing substantial proportions of both French and Latin were identified as *Truly bilingual* ($\mu_F + \mu_L > 0.8$ and $|\mu_F - \mu_L| < 0.2$) or *Massively bilingual* (thresholds 0.7 and 0.3).

	Bilingual		French		Latin		Other	
	Mass.	Truly	Mass.	Truly	Mass.	Truly	Mass.	Ambiguous
Fr	0.5	0.6	35.8	58.2	0.6	0.1	0.9	3.3
Fr, La	3.4	2.9	20.0	5.0	54.6	5.8	0.4	7.9
Fr, La, Ot	1.7	5.1	23.7	6.8	30.5	8.5	8.5	15.3
Fr, Ot	1.9		25.9	13.0	1.9		22.2	35.2
It			0.5		0.5		93.5	5.6
La	1.1	0.6	1.4	0.1	41.6	50.8	0.3	4.1
La, Ot		0.4	1.3	0.4	36.8	19.2	17.1	24.8

Table 4: Language in metadata vs. Language classification over the full content, as percentage.

Manuscripts not fitting any of these criteria were marked as Ambiguous. This hierarchical approach captures both extreme homogeneity and mixed-language phenomena within the corpus, revealing insights that contrast with metadata labels from the metadata providers: Latin and French manuscripts are more often predominantly Latin or bilingual than the metadata suggests (see Table 4). On the other hand, around 90% of manuscripts labelled as “French” or “Latin” are not relabelled.

4.3. Evaluation

To assess the reliability and reusability of the corpus, we carried out a systematic evaluation of transcription quality. Our goal was to quantify recognition accuracy across languages and centuries, and to identify the main sources of error that limit downstream use. The evaluation combines quantitative metrics (character error rate, CER) with a qualitative analysis of common error patterns, thereby providing a comprehensive view of corpus performance.

Sampling strategy To assess the quality of the corpus, we conducted manual post-correction on a sample of 670 manuscripts. For each manuscript, three consecutive lines from a single digitized page were selected and their transcriptions were corrected independently by two expert annotators, one for each language, on eScriptorium (Kiessling et al., 2019). During annotation, each page could be flagged for issues such as bad metadata

(e.g., incorrect century attribution, including 19th-century copies of earlier works, wrong language, or unsuitable material such as early printed books), bad layout analysis (e.g., duplicated lines, incomplete lines, incorrect line direction), or marked for exclusion. Exclusion was applied either when metadata or layout problems made evaluation unreliable, or when the page was deemed too difficult to decipher.

Two groups of manuscripts were defined: one for those labelled as Latin, the other as French. Sampling was stratified by century to ensure temporal coverage, although the distribution of manuscripts across centuries differs between the two languages. When the number of available manuscripts in a given century fell below the target quota, all available manuscripts were included. For centuries with a surplus, the number of manuscripts was capped to maintain balance. As a result, the Latin set includes 52 manuscripts from the 9th century and about 40 per subsequent century. The Old French set contains 12 manuscripts from the 12th century and approximately 80 per century from the 13th onward.

Quantitative Analysis The overall character error rate (CER) across the sample is 9.7%, with Latin manuscripts performing better (8.4%) than French ones (11.05%). Only a small proportion of documents were excluded due to misclassified language, incorrect line detection, or extreme unreadability (11% for Latin, 10.5% for French). Accuracy varies across centuries (Table 5), with French manuscripts of the 12th century showing the widest dispersion of CER values (Figure 1), a trend further explained in the qualitative analysis.

The error distribution reflects well-known challenges in medieval OCR. As in the CATMuS benchmark, spacing errors dominate: 13.8% of all mistakes are insertions, deletions, or substitutions of spaces. Character-level confusions follow predictable palaeographic patterns—‹u/n› substitutions accounts for 3% of errors, ‹i/u› for 2.1%, ‹c/e› for 1.6%, and ‹i/r› for 1.4%—aligning with the inherent ambiguities of medieval scripts.

Qualitative analysis Transcription quality depends primarily on the script type, the layout analysis accuracy, and, to a lesser extent, image quality.

In medieval palaeography, the notion of script type refers broadly to the characteristic style of handwriting used in a given period or region. The most represented script types within the CATMuS dataset, such as the caroline minuscule—an articulated and rounded style common in early medieval Europe—, the textualis —closer to our modern gothic typefaces—, and the hybrida —combining features of textualis and cursiva, a looped script type used especially for administrative or documentary purposes—are generally well recognized by the model. Systematic recognition errors occur naturally when the system encounters script types that are absent from the training data. For example, the presence of the mid-8th-century Beneventan script introduces errors due to the unfamiliar ligatures and letterforms employed.

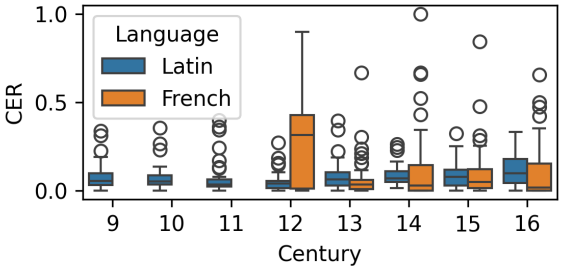

Figure 1: Dispersion of error by manuscript and centuries (CER as %).

	Century	9	10	11	12	13	14	15	16	Total
French	CER				28.6	5.3	14.4	9.7	10.4	
	Support				11	77	68	73	70	299
	Inc. Metadata				6	12	12	7	10	47
	Inc. LA				5	7	17	2	4	35
	Exclusion				1	4	13	8	9	35
Latin	CER	7.6	6.7	8.9	5.6	10.6	8.7	9.5	11.2	
	Support	52	37	36	37	40	37	35	24	298
	Inc. Metadata	0	1	0	1	1	1	2	1	7
	Inc. LA	1	4	8	1	0	3	1	1	19
	Exclusion	4	3	2	3	2	3	5	15	37

Table 5: Average CER (%) per century. The table also reports the number of manuscripts included in CER computation (Support) and page-level annotations of issues: incorrect metadata, incorrect layout analysis, and exclusion from evaluation.

Rapidly executed script type—specifically for the cursiva–and abbreviation rate preset greater challenges. A notable case mixing both is later French account manuscripts–records of financial transactions and inventories—typically written in fast cursiva, which yield particularly high error rates.

Errors linked to layout analysis are mostly due to inaccurate line segmentation, which can produce incomplete or merged lines—for example, when a single predicted baseline spans two or three consecutive lines and prevents correct separation and consequently recognition. Although less critical, low image quality still contributes to recognition errors in faint or small module characters.

5. Example Applications

5.1. Evolution of Layouts

As a first example of a DH application, we analysed the evolution of manuscript page layouts over time. A page in a manuscript is not only a container for

text but also a structured space that organizes information and guides reading (Goody, 1977). Its design evolved gradually, influenced by both copyists and readers (Kwakkel, 2018), yet prior studies were largely qualitative and could not quantify historical patterns across large corpora (Zali, 1999; Grafton, 2012; Vandendorpe, 2021).

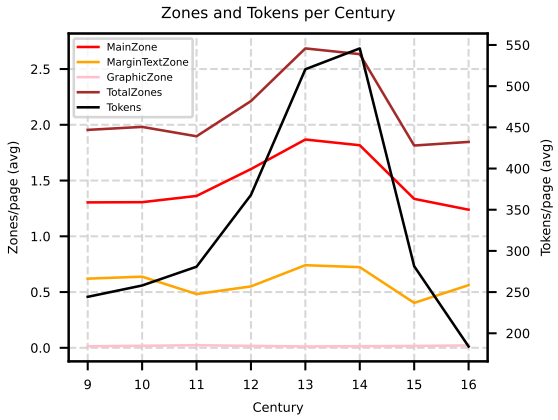

Figure 2: Changes in the average structure of a manuscript page over the centuries (x-axis): smoothed average count of zones (left y-axis, coloured curves) and average number of tokens per page (right y-axis, black curve).

Using our corpus, we quantified layout complexity by analysing `MainZone`'s (text), `MarginTextZone`'s (paratext), and the combination of `DropCapitalZone`'s and `GraphicZone`'s (visual elements), alongside average tokens per page as a proxy for page format (cf. Figure 2).

Three trends emerge: (1) oscillation between mono- and multi-column layouts, with a decline of the former in the 11th century; (2) a phase of increased complexity, marked by more zones and higher token density; (3) simplification towards the Late Middle Ages. These patterns reflect both material constraints, such as manuscript size and cost, and cultural factors, including the influence of scholasticism followed by humanist preferences. This analysis demonstrates how large-scale digitized corpora can enable quantitative studies of manuscript layout evolution, complementing traditional humanities research.

5.2. Scribal Evolution

Complexity appears not only in the layout of the page but also in the form of individual words. For instance, the word *chevalier* ("knight") could be abbreviated as ‹chl̓r› or ‹chr̃›. Such abbreviations entail a cost of decoding time: they are faster to write, but longer to read (Perea et al., 2006).

Fig. 3 shows that Latin manuscripts display a curve similar to that of page layout, peaking

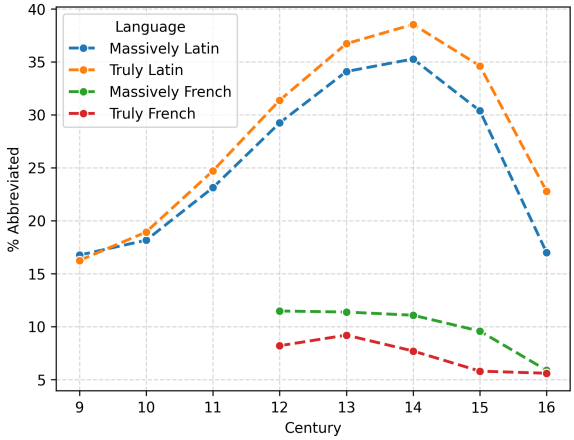

Figure 3: Percentage of abbreviated tokens per century (x-axis) for French and Latin.

in the 14th century and then declining. French manuscripts, in contrast, demonstrate consistently lower percentages, reflecting a different scribal tradition despite—or perhaps because of—their inheritance of Latin abbreviations (Hasenohr, 1998).

When combined with the layout analysis, a general trend emerges: after a phase of maximal complexity around 1300, information presentation gradually simplified toward the end of the period. This pattern may reflect an attempt to optimize the reader's "cognitive load" (Sweller, 1988).

5.3. Contextual Embedding model

Finally, to evaluate the corpus as a resource for NLP, we trained a contextual embedding model. This both demonstrates the corpus's usability for common NLP tasks and establishes a baseline model for future downstream applications.

Training Data Preparation We assembled editorialized Latin and Old French data from multiple sources (see Table 6).From these corpora, we constructed a lowercased vocabulary, excluding single-character tokens and relying on whitespace tokenization. A SentencePiece tokenizer (NFD mode, Kudo and Richardson (2018)) was then trained jointly on the full clean dataset and a subset of 23,000 manuscripts, with a size of 30,002 tokens, including special tokens.

The transformation of the source text into samples followed two distinct strategies. Manuscripts were divided into non-overlapping sequences of five lines. Edited texts, by contrast, were segmented at strong punctuation marks, with a 30% probability of ignoring the punctuation and a minimum sequence length of 20 tokens (as defined by the tokenizer). For validation, 5% of the edited material (173k samples) was held out, with an equal

Corpus (Latin then French)	kTokens
Corpus Corporum (div.) Roelli and Ctibor (2024)	96,995
CEMA Perreaux (2021)	57,601
CAMENA Schibel et al. (1999–2008)	7,916
Assmann and Sahle (1826–)	6,930
DigilibLT Tabacco et al. (2021)	6,674
Perseus Crane et al. (2021b)	6,148
CSEL Crane et al. (2021a)	6,046
CroaLa Jovanović (2024)	5,233
Mirabile Pinelli et al. (2023)	4,655
Bib. Italiana Russo (2023)	4,520
ALIM Stella et al. (2025)	4,284
Add. Texts Clérice (2021)	3,844
Scripta Bauduin et al. (2023)	2,226
Aposcripta Théry (2022)	1,736
ALCAR Stutzmann et al. (2021)	973
Dig. Ptolemaeus Hasse (2023)	880
CGLO Goetz and Gitner	855
SourceEncyMe Draelants and Kuhry (2020)	777
NCA Dees et al. (2006)	3,185
BFM Guillot et al. (2022)	6,293
Fabliaux Pierreville et al. (2025)	120
Geste Camps et al. (2019)	324
DocLing Glessgen et al. (2016)	1,278
TVOF Morcos et al. (2020)	361
OFC Sneddon (1982)	14
Total	229,870

Table 6: Normalized corpus for ModernBert.

number of samples withheld from the manuscript set, yielding a total validation set of 347k samples.

The manuscript training data was further filtered using the vocabulary list derived above. Any manuscript sample with less than 30% vocabulary overlap was discarded, as these typically corresponded to heavily abbreviated or noisy ATR outputs. A manual check on 400 edited texts yielded an average overlap of 57%, supporting this threshold. Filtering resulted in removing 20% of manuscript samples, leaving 40.5 million usable sequences, supplemented by 3.3 million sequences from edited corpora. Finally, to balance the training signal, clean editorial samples were upsampled following the approach of Bamman and Burns (2020), ensuring that each epoch presented equal proportions of edited and ATR-derived data.

Experimental Set-up We trained a ModernBERT model using the HuggingFace library with default parameters, except for the vocabulary size and the mask replacement probability, which was changed to 98% to account for the inherent noise of the data. The model was trained for 10 epochs. Hyperparameter tuning and adjustments added the equivalent of 3 additional days of computation.

Method We selected two digitized edited texts and two manuscripts that were not part of the previously described datasets. From each sentence, we extracted 100 random segments. In each segment, a single token was randomly masked, and the model's predictions were evaluated against the top-1 and top-5 outputs. Expert annotators reviewed the results, taking into account potential historical spelling variants, which extended the scoring to include variant spellings.

File	Top 1	Top 5	Top 1 Var.	Top 5 Var.
Fro (Edited)	54.4	68.9	58.3	70.9
Fro (ATR)	45.6	65.0	56.3	68.9
Lat (Edited)	53.4	72.8	55.3	73.8
Lat (ATR)	33.3	52.9	40.2	53.9

Table 7: Masked filling accuracy over 4 new texts.

Results The model demonstrated a strong ability to provide plausible token predictions, as shown in Table 7. Old French benefited the most from considering spelling variants (+3.9 for Top-1 in edited texts, +10.7 in ATR output). Interestingly, the model sometimes proposed abbreviated spelling variants in the ATR outputs for both Latin and French, but never for the edited texts. In at least one instance, the model did not reproduce the masked token exactly but produced the correct reading, since the ATR itself contained an error at that position. For Latin, Top-1 predictions occasionally suggested a different tense or mood of the same verb, which remained grammatically acceptable in context, yet was counted as faulty.

6. Conclusion

In this paper, we introduced a large-scale corpus of medieval manuscripts obtained through automatic text recognition, covering a little under 2.5 billion tokens in Latin and Old French. The resource is openly distributed in ALTO, TEI-XML, plain-text and JSON formats enabling research in computational philology, corpus lingusistics, NLP, and DH. By providing raw, non-normalized text enriched with structural information, the corpus establishes a new foundation for large-scale experimentation with medieval textual data, from language modelling to the analysis of manuscript layouts.

As future work, we aim to expand the coverage of the corpus both linguistically and geographically. In particular, we plan to integrate Spanish materials from digitized medieval archives, as well as additional Latin and French sources from underrepresented repositories. These extensions will contribute to a broader multilingual resource, enhancing the robustness of NLP models trained on medieval data and opening new perspectives for comparative studies of manuscript culture across Europe. Ultimately, our goal is to establish a scalable, multilingual, and sustainable corpus that will serve the research community for years to come.

7. Acknowledgments

We would like to thank Régis Robineau and the entire Biblissima+ team for their generous support in designing and running the queries that enabled access to the digitized manuscripts used in this work. This research was conducted in the context of the COLaF and the ParamHTRs projects, respectively, funded by Inria and by the BnF Datalab and the TGIR Huma-Num. Their support and infrastructure have been essential in making this large-scale corpus possible.

8. Bibliographical References

Sergio Torres Aguilar. 2024. Handwritten text recognition for historical documents using visual language models and gans.

Sergio Torres Aguilar. 2025. Tridis: A comprehensive medieval and early modern corpus for htr and ner. *arXiv preprint arXiv:2503.22714*.

David Bamman and Patrick J Burns. 2020. Latin bert: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.

Natalia Bottaioli, Solène Tarride, Jérémy Anger, Seginus Mowlavi, Marina Gardella, Antoine Tadros, Gabriele Facciolo, Rafael Grompone von Gioi, Christopher Kermorvant, Jean-Michel Morel, et al. 2024. Normalized vs diplomatic annotation: A case study of automatic information extraction from handwritten uruguayan birth certificates. In *International Conference on Document Analysis and Recognition*, pages 40–54. Springer.

Eltjo Buringh. 2010. *Medieval Manuscript Production in the Latin West*.

Roberto Busa. 1980. The annals of humanities computing: The index thomisticus. *Computers and the Humanities*, pages 83–90.

Centre Traditio Litterarum Occidentalium. 2025. Library of latin texts (online). Database, Brepols Publishers.

Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2017. Icdar2017 competition on layout analysis for challenging medieval manuscripts. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 1371–1376.

Thibault Clérice, Ariane Pinche, Malamatenia Vlachou-Efstathiou, Alix Chagué, Jean-Baptiste Camps, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, Simon Gabay, Patricia O'connor, Wouter Haverals, Mike Kestemont, Caroline Vandyck, and Benjamin Kiessling. 2024. CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond. In *Lecture Notes in Computer Science*, volume 14806 of *Lecture Notes in Computer Science*, pages 174–194, Athens, Greece.

Giorgia Crosilla, Lukas Klic, and Giovanni Colavizza. 2025. Benchmarking large language models for handwritten text recognition.

Anthonij Dees. 1985. Dialectes et scriptae à l'époque de l'ancien français. *Revue de Linguistique Romane Lyon*, 49(193-194):87–117.

Joseph Denooz. 1996. La banque de données du laboratoire d'analyse statistiques des langues anciennes (lasla). *Le Médiéviste et l'ordinateur*, 33(1):14–20.

Daniel Hernandez Diaz, Siyang Qin, Reeve Ingle, Yasuhisa Fujii, and Alessandro Bissacco. 2021. Rethinking text line recognition models. *arXiv preprint arXiv:2104.07787*.

Simon Gabay and Thibault Clérice. 2024. The birth of French orthography. A computational analysis of French spelling systems in diachrony. In *CHR2024 – Computational Humanities Research Conference*, Aahrus, Denmark.

Simon Gabay, Ariane Pinche, Kelly Christensen, and Jean-Baptiste Camps. 2024. Segmonto: a controlled vocabulary to describe and process digital facsimiles. *Journal of Data Mining and Digital Humanities*.

Martin-Dietrich Glessgen. 2017. Les «documents linguistiques» de la france. histoire, présent et perspectives d'un projet centenaire. *Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres*, 161(3):1261–1292.

Jack Goody. 1977. *The Domestication of the Savage Mind*. Cambridge University Press, Cambridge.

Anthony Grafton. 2012. *La Page de l'Antiquité à l'ère du numérique. Histoire, usages, esthétiques*. Hazan / Musée du Louvre, Paris.

Tobias Grüning, Gundram Leifert, Tobias Strauß, Johannes Michael, and Roger Labahn. 2018. Read-bad: A new dataset and evaluation scheme for baseline detection in archival documents. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 351–356.

Céline Guillot, Serge Heiden, and Alexei Lavrentiev. 2018. Base de français médiéval: une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques. Revue de Linguistique française diachronique*, (7):168–184.

Geneviève Hasenohr. 1998. Abréviations et frontières de mots. *Langue française*, 119:24–29.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Rahima Khanam and Muhammad Hussain. 2024. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

Benjamin Kiessling. 2022. The Kraken OCR system.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. escriptorium: An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Erik Kwakkel. 2018. The architecture of the medieval page. *medievalbooks*.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102.

William Mattingly. 2025. Biglam catmus medieval manuscript yolov11x.

Matthew Thomas Miller, Maxim G. Romanov, and Sarah Bowen Savant. 2018. Digitizing the textual heritage of the premodern islamicate world: Principles and plans. *International Journal of Middle East Studies*, 50(1):103–109.

Véronique Montémont. 2020. De frantext 1 à frantext 2: la cure de jouvence d'une vieille dame. *La lexicographie informatisée: les vocabulaires nationaux dans un contexte européen*, page 41.

Yves Charles Morin. 2006. Histoire du corpus d'Amsterdam: le traitement des données dialectales. *Le Nouveau Corpus d'Amsterdam, Actes de l'atelier de Lauterbad*.

Emmanuelle Morlock, Anne-Marie Turcan-Verkerk, Régis Robineau, Eduard Frunzeanu, and Kevin Bois. 2025. Unlocking the Past: The Biblissima Portal, a Gateway to Ancient Written Heritage in the Digital Age. In *DARIAH Annual Event 2025 - The Past*, Göttingen (DE), Germany. Dariah.

Silvio J. F. Oliveira, Benoit Seguin, and Frederic Kaplan. 2018. dhsegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 7–12.

Manuel Perea, Joana Acha, and Manuel Carreiras. 2006. Eye movements when reading text messaging (txt msgng). *Quarterly Journal of Experimental Psychology*, 62:1560–1567.

Ariane Pinche. 2022. Guide de transcription pour les manuscrits du Xe au XVe siècle.

Ariane Pinche, Thibault Clérice, Malamatenia Vlachou-Efstathiou, Alix Chagué, Jean-Baptiste Camps, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, Simon Gabay, and Genaro Ferrante. 2025. Catmus medieval.

Philipp Roelli and Jan Ctibor. 2024. A new version of corpus corporum, the latin full-text database and tool. *Bulletin du Cange (Archivium latinitatis medii aevi)*, 80:251–266.

Achim Stein and Pierre Kunstmann. 2006. Le nouveau corpus d'amsterdam: Corpus informatique de textes littéraires d'ancien français (ca. 1150-1350). Online resource, ATILF / Université de Stuttgart, Institut für Linguistik/Romanistik.

Daniel Stoekl Ben Ezra, Luigi Bambaci, Nachum Dershowitz, Benjamin Kiessling, and Avi Shmidman. 2025. Large scale computational analysis of historical manuscripts: The midrash project and its applicability to other cultures. In *Proceedings of the 13th Conference of the Japanese Association for Digital Humanities*. "Leveraging AI and Digital Humanities for Sustainable Infrastructure".

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12:257–285.

Solène Tarride, Yoann Schneider, Marie Generali-Lince, Mélodie Boillet, Bastien Abadie, and Christopher Kermorvant. 2024. Improving automatic text recognition with language models in the pylaia open-source library. In *International Conference on Document Analysis and Recognition*, pages 387–404. Springer.

Yanco Amor Torterolo-Orta, Jaione Macicior-Mitxelena, Marina Miguez-Lamanuzzi, and Ana García-Serrano. 2025. Transcribing spanish texts from the past: Experiments with transkribus, tesseract and granite.

Christian Vandendorpe. 2021. The page in the western tradition: From the clay tablet to the digital format. *Architectures of the Book*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Lianwen Yang, Lianwen Jin, Xing Huang, and Ke Deng. 2017. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 406–413.

Anne Zali. 1999. *L'aventure des écritures III : la page*. Bibliothèque nationale de France, Paris.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception.

9. Language Resource References

Assmann, B. and Sahle, Patrick. 1826–. *Monumenta Germaniae Historica*.

Bauduin, Pierre and Buard, Pierre-Yves and Goloubkoff, Anne and Mancel, Émeline and Fournier-Fujimoto, Tamiko and Buquet, Thierry and Brossard, Stéphane and Trohel, Benoît and Combalbert, Grégory and Bisson, Marie and Berthelot, Clémentine and Dubois, Adrien and Bloche, Michaël. 2023. *Scripta*.

Jean-Baptiste Camps and Alice Cochet and Lucence Ing and Elena Albarran. 2019. *Geste: un corpus de chansons de geste, 2016-…*. Zenodo.

Clérice, Thibault. 2021. *Corpora of rare texts, Lasciva Roma*.

Crane, Gregory Ralph and Babeu, Alison and Cerrato, Lisa and Munson, Matthew and Dee, Stella and Gessner, Annette and Robertson, Bruce and Franzini, Greta and Stoyanova, Simona and Selle, Tabea and Springmann, Uwe and Clérice, Thibault. 2021a. *Corpus Scriptorum Ecclesiasticorum Latinorum digitized by Open Greek And Latin*.

Crane, Gregory Ralph and Mylonas, Elli and Smith, Neel and Almas, Bridget and Babeu, Alison and Berra, Aurélien and Bonhomme, Marie-Laurence and Brooks, David and Cerrato, Lisa and Clérice, Thibault and Dee, Stella and Ehwald, R. and Frering, Léa and Gessner, Annette and Harrington, J. Matthew and Himes, Zach and Jovanovic, Neven and Konieczny, Michael and Mahoney, Anne and Merrill, William and Mimno, David and Mueller, Lucian and Munson, Matthew and Singhal, Rashmi and Smith, David and Taounza-Jeminet, Alyx and Weaver, Gabriel A. and Wulfman, C. 2021b. *Perseus Canonical Latin Literature*.

Anthonij Dees and Achim Stein and Pierre Kunstmann and Martin Gleßgen and Giles Souvay. 2006. *Nouveau Corpus d'Amsterdam*.

Draelants, Isabelle and Kuhry, Emmanuelle. 2020. *Sources des Encyclopédies Médiévale*.

Glessgen, Martin-Dietrich and Carles, Hélène and Duval, Frédéric and Videsott, Paul. 2016. *Documents linguistiques galloromans*.

Goetz, Georg and Gitner, Adam, et al. *CGLO*. 23-12.

Guillot, Céline and Heiden, Serge and Lavrentiev, Alexei. 2022. *Base de français médiéval*.

Hasse, Dag Nikolaus. 2023. *Ptolemaeus Arabus et Latinus*.

Neven Jovanović. 2024. *Croatiae Auctores Latini Textus*. Zenodo.

Hannah Morcos and Simon Gaunt and Maria Teresa Rachetta and Henry Ravenhall and Simone Ventura and Luca Barbieri and Paul Caton and Geoffroy Noël and Ginestra Ferraro and Natasha Romanova. 2020. *The Histoire ancienne jusqu'à César: A Digital Edition (The Values of French)*.

Perreaux, Nicolas. 2021. *Cartae Europae Medii Aevi (CEMA)*.

Corinne Pierreville and Jules Nuguet and Ariane Pinche and Alexei Lavrentiev. 2025. *Projet Fabliaux*.

Pinelli, Luca and Ctibor, Jan and Roelli, Philipp and Degl'Innocenti, Antonella and Gamberini, Roberto and Carlamaria Crespi, Serena and Fizzarotti, Luisa and Montepaone, Olivia and Vangone, Laura and Santarelli, Riccardo. 2023. *Mirabile Digital Archive*.

Roelli, Philipp and Ctibor, Jan. 2024. *Corpus Corporum*.

Russo, Emilio. 2023. *Biblioteca Italiana*.

Schibel, Wolfgang and Kredel, Heinz and Niehl, Rüdiger. 1999–2008. *CAMENA: Latin Texts of Early Modern Europe*. DFG-funded project; online access.

Clive R. Sneddon. 1982. *Old French corpus*. Literary and Linguistic Data Service.

Francesco Stella and Edoardo D'Angelo and Giorgio Di Maria and Marina Buzzoni and Fulvio Delle Donne and Roberto Rosselli Del Turco. 2025. *ALIM: Archivio della Latinità Italiana del Medioevo*. Accessed: 2025-09-27.

Stutzmann, Dominique and Torres Aguilar, Sergio and Chaffenet, Paul. 2021. *HOME-Alcar: Aligned and Annotated Cartularies*. Zenodo.

Tabacco, Raffaella and Lana, Maurizio and Balossino, Michele and Bessi, Giancarlo and Bognini, Filippo and Buffa, Martina and Caso, Daniela and Cattaneo, Gianmario and Ciotti, Fabio and Ciusani, Mauro and Colombo, Piero and Cuzzotti, Claudia and del Core, Vincenzo and Della Calce, Elisa and Denicola, Luciano and Digirolamo, Letizia and Ferrandi, Etienne and Ferroni, Manuela and Fontana, Davide and Guiglia, Valeria and Loberti , Martina and Lucciano, Melanie and Maconi, Ludovica and Malaspina, Ermanno and Manuela, Ferroni and Marini, Alessia and Maronet, Léa and Massano, Federico and Mazzucco, Clementina and Mellano, Anastasia and Miglietta, Chiara and Mollea, Simone and Mosca, Laura and Musso, Simona and Naso, Manuela and Paniagua, David and Poncina, Fabio and Ramires, Giuseppe and Rinaldi, Valentina and Rosso, Nadia and Rota, Simona and Rozzi, Stefano and Rugnone, Elisa and Senore, Corinna and Stok, Fabio and Strona, Beatrice and Verny, Romain and Vittoria Martino, Maria. 2021. *Digital Library of Late-Antique Latin Texts*.

Théry, Julien. 2022. *APOSCRIPTA database. Unified Corpus of Papal Letters*. Zenodo.