CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond

Thibault Clérice, Ariane Pinche, Malamatenia Vlachou-Efstathiou, Alix Chagué, Jean-Baptiste Camps, Matthias Gille-Levenson, Olivier Brisville-Fertin, Franz Fischer, Michaels Gervers, Agnès Boutreux, et al.

CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond

Thibault Clérice[1 [0000−0003−1852−9204]], Ariane Pinche[2 [0000−0002−7843−5050]], Malamatenia Vlachou-Efstathiou[3,14 [0000−0002−9397−356X]], Alix Chagué[1,4,14 [0000−0002−0136−4434]], Jean-Baptiste Camps[5,6 [0000−0003−0385−7037]], Matthias Gille Levenson[5,2,13 [0000−0001−9488−5986]], Olivier Brisville-Fertin[2,13 [0000−0001−7809−3890]], Federico Boschetti[7,8 [0000−0002−7810−7735]], Franz Fischer[8], Michael Gervers[9 [0000−0003−2381−7964]], Agnès Boutreux[9], Avery Manton[9], Simon Gabay[10 [0000−0001−9094−4475]], Patricia O'Connor[5 [0000−0001−5880−4704]], Wouter Haverals[11 [0000−0002−5687−6787]], Mike Kestemont[12 [0000−0003−3590−693X]], Caroline Vandyck[12 [0009−0006−9995−1325]], and Benjamin Kiessling[13 [0000−0001−9543−7827]]

[1] ALMAnaCH - Automatic Language Modelling and Analysis & Computational Humanities, Inria, Paris, France
[2] CIHAM–UMR 5648, CNRS, Lyon, France
[3] IRHT - Institut de Recherche et d'Histoire des Textes, Paris, France
[4] UdeM - Université de Montréal, Montréal, Canada
[5] CJM - Centre Jean Mabillon, Paris, France
[6] ÉNC - École nationale des chartes, Paris, France
[7] ILC-CNR, Pisa, Italy
[8] VeDPH - Venice Centre for Digital and Public Humanities, Ca'Foscari, Venice, Italy
[9] UToronto - Department of History, University of Toronto, Canada
[10] UNIGE - Université de Genève, Switzerland
[11] Princeton University, Princeton NJ 08544, USA
[12] Antwerp University, Belgium
[13] École Normale Supérieure de Lyon, France
[14] EPHE, France

Abstract. The surge in digitisation initiatives by Cultural Heritage institutions has facilitated online accessibility to numerous historical manuscripts. However, a substantial portion of these documents exists solely as images, lacking machine-readable text. Handwritten Text Recognition (HTR) has emerged as a crucial tool for converting these images into machine-readable formats, enabling researchers and scholars to analyse vast collections efficiently. Despite significant technological progress, establishing consistent ground truth across projects for HTR tasks, particularly for complex and heterogeneous historical sources like medieval manuscripts in Latin scripts (8^{th}-15^{th} century CE), remains nonetheless challenging.

We introduce the Consistent Approaches to Transcribing Manuscripts (CATMuS) dataset for medieval manuscripts, which offers (1) a uniform

framework for annotation practices for medieval manuscripts, a benchmarking environment (2) for evaluating automatic text recognition models across multiple dimensions thanks to rich metadata (century of production, language, genre, script, etc.), (3) for other tasks (such as script classification or dating approaches), (4) and finally for exploratory work pertaining to computer vision and digital paleography around line-based tasks, such as generative approaches.
Developed through collaboration among various institutions and projects, CATMuS provides an inter-compatible dataset spanning more than 200 manuscripts and *incunabula* in 10 different languages, comprising over 160,000 lines of text and 5 million characters spanning from the 8$^{\text{th}}$ century to the 16$^{\text{th}}$. The dataset's consistency in transcription approaches aims to mitigate challenges arising from the diversity in standards for medieval manuscript transcriptions, providing a comprehensive benchmark for evaluating HTR models on historical sources.

Keywords: Historical sources · medieval manuscripts · Latin scripts · benchmarking dataset · multilingual · handwritten text recognition.

1 Introduction

Cultural heritage institutions, propelled by the digitization wave of the last two decades, have made tens of thousands of manuscripts accessible online. However, these invaluable historical documents predominantly exist in image form, lacking machine-readable text. The surge in interest in automatic handwriting recognition for historical documents, as noted by Fischer et al. in 2011 [16], has only intensified, particularly with the emergence of all-in-one platforms such as Transkribus [29] and eScriptorium [34]. Offline Handwritten Text Recognition (hereafter HTR) has become an indispensable tool for philologists, historians, linguists, librarians, and archivists, providing large-scale transcriptions of documents that would otherwise remain largely untouched for research purposes [8].

While the advancements in HTR technology are commendable, producing a meaningful and diverse benchmark dataset remains a significant challenge. This challenge is particularly pronounced for manuscripts in the Latin script dating from the medieval millennium (500-1500 CE), where a substantial volume of documents awaits transcription. The precise number of such manuscripts is unknown, primarily due to the incomplete nature or overall absence of catalogues for every public or private collection. However, Buringh [7, p.,99] estimates that approximately 1,300,000 manuscripts from the Latin West, spanning the 1$^{\text{st}}$ to the 19$^{\text{th}}$ century CE, currently survive.

Simultaneously, medieval manuscript sources provide a distinctive playground for computer vision, particularly in the field of HTR, as the nature of the data presents a unique set of challenges and opportunities. Specifically, the dynamic evolution of language during the medieval millennium, manifested in linguistic variations across geographic regions and the absence of strict orthographic rules, along with variations in scribal practices at both collective and individual levels compared to their modern counterparts, establishes a rich and demanding

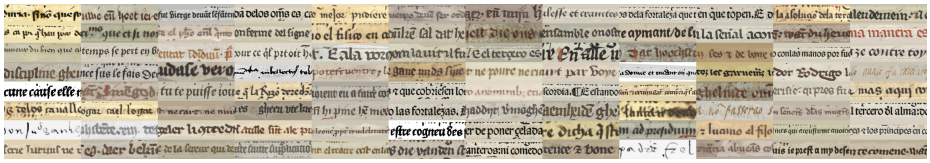

Fig. 1: 96 random sample lines from the approximately 160,000 lines found in the dataset, written in different languages and scripts.

Fig. 2: Several transcriptions are possible for these two images. On the left, ‹de la uirtut› and ‹dela ꝟırtut› are both acceptable transcriptions, depending on the annotation guidelines. Likewise, on the right, ‹uirũ principatu› could be normalized as ‹virum principatu›.

testing ground for exploring HTR capabilities. In this context, systems can be pushed to their limits (see Figure 1).

Due to conflicting or established traditions, transcription standards for medieval manuscripts exhibit significant diversity, thereby posing a considerable hurdle in creating a uniform framework for evaluating HTR models. Existing datasets adhere to a variety of project-specific practices, leading to compatibility issues in terms of data aggregation[15]. While some datasets include transcriptions according to printed edition standards, with resolved abbreviations [59], others choose to maintain all parallel graphic variants for specific "characters", such as ‹s›/‹ſ› [24]. In contrast, our approach shifts the focus to the semantic weight of written signs, "merging" variant representations of the same "character" (*cf.* Figure 2), including diacritics such as macrons and tildes into one representative sign.

To address this visual diversity while maintaining common guidelines, we introduce the Consistent Approaches to Transcribing Manuscripts (CATMuS) and its medieval dataset, specifically tailored for evaluating HTR architectures. Collaboration among a dozen projects has resulted in a large-scale cohesive dataset comprising nearly 200 manuscripts or early printed documents in 10 different languages, containing over 160,000 lines. Presented in Parquet format, the dataset includes rich and historically relevant metadata such as century, language, script class, and genre.

While HTR and other discriminative methods, such as writer identification, script, and date classification, are the most prominent applications of our dataset, it could also cater to emerging techniques like Condition-Adaptive HTR [3], computer-assisted palaeographical analysis [51], as well as Handwriting Text Generation (HTG). For instance, through the utilization of Generative Adver-

[15] A similar tendency has been observed in the treatment of digital historical corpora [21].

sarial Networks (GANs) for the creation of synthetic data [61,31,19,30], we can leverage our extensive labeled dataset to generate under-represented features such as specific characters, script types, or period-related styles.

Before the detailed presentation of the related work, our contributions can be summarized as follows:

- Release of a large consistently annotated dataset in Latin script spanning over 9 centuries as text-image pairs in Parquet formats, enabling both image-to-text and text-to-image approaches given the scale of the data.
- Metadata that enable experimentation for the previously mentioned tasks as well as different classification tasks, including script and century classification.
- A thorough description of our transcription practice and an analysis of the datasets.
- Benchmarking results for HTR.

Hereafter, this paper is organized into four distinct sections as follows: First, Section 2 provides an overview of parallel datasets, highlighting their annotation practices, metadata, and task specificities. Section 3 presents the CATMuS dataset, detailing general and specific characteristics as well as our transcription guidelines rationale, along with statistics. Finally, Section 4 presents standard benchmarking splits and results for two baseline models, followed by a brief discussion of the results.

2 Related Works

In terms of available data, numerous medium and large-size datasets have been released for HTR or specific tasks related to historical document analysis [37]. Some of these datasets have been utilized in HTR competitions organized within the framework of the ICDAR and ICFHR conferences. Only the most relevant datasets to ours, in terms of format (text-line annotation pairs), time-span, metadata, and tasks, are presented below to facilitate comparison.

2.1 Benchmarking Datasets for Historical Handwritten Text Recognition

In the domain of Historical HTR, various benchmarking datasets have gained attention, notably those focused on restricted periods such as the Middle Ages. Three main datasets serve as widely used benchmarks in HTR competitions, providing a foundational basis for evaluating model performance and advancements: the **Perzival Database** [18] and the **Saint Gall Database** [17] from IAM-HistDB, and the ICFHR-2016 READ Dataset [60,48]. The first two address each a single document written by a single hand: Perzival comes from a manuscript from the 13^{th} century in Middle High German, and the second one

Latin manuscripts from the 9th century in Carolingian script. Despite being issued from the same "Historical Database," *PerzivalDB* and Saint Gall DB are incompatible when it comes to their transcription practices: the first one preserves parallel variants of letters – such as ‹s› and ‹ſ› – while the second one normalizes forms (such as ‹virum›), drops abbreviating diacritics such as macrons, and finally resolves specific signs such as ‹&› into "et." The **ICFHR-2016 READ** (Recognition and Enrichment of Archival Documents) *dataset* consists of a subset of documents from the "Ratsprotokolle collection" in Early Modern German, composed of approximately 30,000 pages of council meeting minutes from 1470 to 1805 written by several hands, which includes the very late Middle Ages with a much larger coverage of modern times.

2.2 Related Datasets for Historical Handwritten Text Recognition

Expanding beyond conventional benchmarks, there are several language-specific and period-specific datasets available in Latin characters, the details of which we elaborate on below. Despite their richness, these datasets may exhibit limitations for generalist approaches, including language specificity, temporal constraints, and project-specific variations in annotation practices.

Starting from French and Latin, the **ORIFLAMMS** Consortium macro-dataset, composed of several different projects over a substantial interval of time, includes **ECMEN**[53], a subset of the dated and datable manuscripts from the Bibliothèque Nationale de France, **PsautierIMS**[54], transcriptions of the text of the biblical Psalm 101, and **Fontenay** [55], containing charters from the Cistercian abbey Fontenay in Burgundy, providing a unique set of data. These datasets offer a diverse array of regional scripts from 779 manuscripts, ranging from the 12th to the 14th century, and offer meticulous metadata regarding provenance and date of production. They provide a mix of aligned preexisting normalized editions (without abbreviations) and graphemic transcriptions (including abbreviations and their expansion, but not regarding capital letters, punctuation, and spacing[16]).

The most extensive source for documentary scripts[17] is **HIMANIS**(Historical MAnuscript Indexing for user-controlled Search) [57], which contains registers produced by the French Royal Chancery between 1302 and 1483. Ground truth was established by aligning digitized images line by line with the partial semi-diplomatic edition of the text by Paul Guérin. In the same vein, but going a bit further, the **e-NDP** (e-Notre Dame de Paris) [12], following an "edition-based" transcription, includes a total of 500 pages from 26 registers dating from 1326 to 1504, and the **HOME-Alcar** project [56] includes 3090 acts from 17 French cartularies dating from the 12th to the 14th centuries respectively. The former is transcribed with edition norms, while the latter is aligned with scholarly editions, and both datasets include entities for important places and names.

[16] Note that all of the ORIFLAMMS datasets underwent minimal normalization when combined with other datasets by another research team [9].

[17] Documentary scripts in the Middle Ages are developed alongside Bookscripts, the latter used for literary manuscripts, and they are *Cursiva*-adjacent scripts.

Still within the Latin alphabet inventory but for manuscripts in modern scripts[18], for Castilian, the most extensive parallel dataset, **RODRIGO** [49], concerns one 1545 manuscript in old Castilian by one writer. A smaller set of initial name indexes, the **Esposalles**/INDEX[46] subset, includes 29 pages in Catalan by a single writer between 1491 and 1495. The transcription guidelines are unique for these two projects, as its base is imitative in terms of punctuation, capitals, and special signs, but with the addition of the symbol $ after every word cut in a page break, characters for missing natural blank spaces between successive words that are indicated by the symbol ‹⌣›, and super/subscript elements that follow LaTeX-like notation as `_{sub}` and `^{super}`, respectively (e.g., `q^{i}`er for superscript "i" in the word "quier"). For Italian, the **LAM** (Ludovico Antonio Muratori) dataset [11] stands out as the most extensive for the 17$^{\text{th}}$ century concerned with 72 manuscript samples from one hand over 60 years, with their diplomatic annotations.

All solely edition-based datasets present normalized punctuation, spacing, and capitals, as well as resolved abbreviation (see Table 1). The issue of transcribing texts with resolved abbreviation (following the edition-based paradigm) has been explored in several recent articles [10,47]. Torres et al. [58] acknowledges that abbreviations[19] are part of the most common errors of their models, which could only be resolved when fine-tuning on a specific collection. Moreover, each of these papers, including the latter, fails to evaluate the generalizability of the models in cross-genre, diachronic, and multilingual settings, as they at most offer a bilingual setting but in very repetitive genres in only a few manuscripts.

2.3 Latin Palaeography Task-Related Datasets

Further delving into specific tasks related to Latin Palaeography, datasets such as **CLaMM** (Classification of Latin Medieval Manuscripts)[13], and **MPS** (Medieval Paleographical Scale) focus on Latin script classification and dating, two tasks closely intertwined for historical analysis. CLaMM, designed for the classification of 12 book script types, encompasses Latin medieval manuscripts from the 5$^{\text{th}}$ to the 16$^{\text{th}}$ century. Originally employed in the ICFHR 2016 Competition for Script Classification, CLaMM comprises 10,800 images.

Similarly, the **MPS** dataset [28], utilized in ICFHR 2014 for manuscript dating, focuses on handwriting-based dating within a corpus of 1,706 charters originating from the Middle Dutch language area, spanning from 1300 to 1550. For *incunabula* and early print[20], the "Dataset of pages from early printed books with multiple font groups" [50], includes 35,623 images of pages written in 12

[18] We consider modern scripts the scripts developed after the Humanistic script, namely personal hands, like "Secretary" and Italic hands that resemble modern handwriting.

[19] They provide the example of short expansions ("no[-bis]"), heavy contractions ("m[a]g[is]t[er]"); declension and syntax-dependent expansion ("Par." which can be "Par[is]", "Par[isiensis]" or "Par[isiense]"), among others.

[20] The term *incunabula* refers to the books printed before the year 1501 using movable type.

Table 1: Overview of parallel datasets. *Abbr. Res.* identifies datasets with resolved abbreviations. *Char. Var.* identifies the presence of multiple classes for single letters (e.g. ‹s›/‹ſ›). *Modern.* stands for edition-like modernization (e.g. spelling). Mss stands for Manuscripts, EP for Early Prints, SC/FC/DC for Script, Font, and Date Classification, l. for lines, d. for documents, i. for images.

Dataset	Type	Century	Language	Script type	Task	Abbr. Res.	Char. Var.	Modern.	Quantity
IAMHistDB/St. Gall	Mss	9	Latin	Bookscript	HTR				1,410 l.
IAMHistDB/Perzival	Mss	13	German	Bookscript	HTR	✓	✓		4,477 l.
ICFHR 2016-READ	Mss	15-19	Ger.	Modern	HTR				10,550 l.
ORIFLAMMS	Mss	12-14	French, Lat.	Book., Docum.	HTR	✓	✓	✓	120,111 l.
HIMANIS	Mss	14-15	Fr., Lat.	Documentary	HTR	✓	✓	✓	23,112 l.
HOME - Alcar	Mss	12-14	Fr., Lat.	Bookscripts	HTR, NER	✓		✓	74806 l.
e-NDP	Mss	14-16	Fr., Lat.	Documentary	HTR	✓		✓	33,735 l.
RODRIGO	Mss	16	Spanish	Bookscript	HTR				20,357 l.
Esposalles/INDEX	Mss	15	Catalan	Modern	HTR				1,563 l.
LAM	Mss	17-18	Italian	Modern	HTR, DC				25,823 l.
MPS	Mss	13-16	Dutch	Documentary	DC				1,706 d.
CLAMM	Mss	5-15	Lat.	Bookscripts	SC, DC				10,800 i.
MFG[22]	EP	15-17	Multi.	Fonts	FC				35,623 i.
CATMuS (Ours)	Mss, EP	8-16	Multi.	Bookscripts Documentary Fonts	HTR, SC, DC				165,347 l.

fonts, including Greek and Hebrew, occasionally multilabel lines, for font and multilabel classification methods.[21]

3 The CATMuS-Medieval Dataset

3.1 Extraction and Annotation Workflow

Data Collection and Preparation The CATMuS Medieval dataset is a macro-dataset, derived from various sources (cf. Table 2): (1) datasets originally produced following guideline norms, (2) datasets easily converted due to shared or compatible transcription practices, and (3) manually corrected datasets. Its creation unfolded through three distinct project phases; in the initial phase, the CREMMA and HTRomance projects were instrumental, with their members leading the formulation of primary transcription guidelines. These guidelines primarily focused on literary manuscripts written in book scripts, spanning Old French, Latin, Castilian, and Italian. The second phase witnessed the involvement of the DEEDS project, focusing on Latin cartularies. Finally, the third phase saw the adoption of our guidelines by the team working on the Monastery of Herne and research on Middle English.

Segmentation and Transcription In our dataset, eScriptorium[34] serves as the primary software used for annotation across projects, except for Mid-

[21] This dataset was used alongside CLaMM and extra provenance annotations for script classification, document dating, and localization in ICDAR2021.

Table 2: Reviewers are included in annotators. "Many" represents situations where the number of annotators is superior to three or unknown.

Repository	Type	Annotators	Documents	Lines	Characters
Carthusian Monastery Of Herne	Manuscripts	Many	18	47,322	1,560,687
Gallicorpora [43,20,42]	Manuscripts, Early prints, Incunabula	5	28	16,515	605,938
CREMMA[41,14]	Manuscripts	4	35	29,722	1,048,144
HTRomance[35,1,23,5]	Manuscripts	8	52	16,574	611,244
PSL-Chartes-HTR-Students/HN2021-Boccace[63]	Incunabula	Many	2	3647	126,677
PSL-Chartes-HTR-Students/decameron-fr[4]	Manuscripts	Many	1	751	23,278
Towards General Castilian HTR[22]	Manuscripts	1	29	29,043	1,003,888
adhoc/matthias	Manuscripts	1	3	253	10,728
adhoc/patricia	Manuscripts	1	1	586	18,396
ciham/fabliaux[44]	Manuscripts	1	5	2070	54,245
ciham/liber[2]	Manuscripts	Many	4	3788	159,799
DEEDS Project	Manuscripts	1	9	2327	153,067
malamatenia/Eutyches[62]	Manuscripts	1	4	12,309	372,528
rescribe/carolineminuscule-groundtruth[27]	Manuscripts	Many	17	440	19,890
Total			208	165,347	5,768,509

dle Dutch data transcribed using Transkribus[29]. The common transcription practice involves employing a line and region segmentation model such as BLLA on eScriptorium [33], with corrections typically including cutting non-continuous lines or expanding them to ensure correct coverage of the text. Following Pinche's experiment regarding the compatibility of Transkribus and eScriptorium data [40], we repolygonized the data from Transkribus and post-corrected the inconsistencies. After segmentation, in the majority of our projects, the ground-truth results in the post-correction of an automatic transcription (see Section 3.2).

Extraction and Post-Processing All lines are extracted from images with XML files containing segmentation coordinates (baselines and masks) and their corresponding text label. The extraction process is twofold: first, the minimal bounding box around the bounding polygonal mask is used to crop the masks, followed by filling the area outside the mask with a proxy value approximating the color of the writing surface. For colored images, we use the the median value of pixels inside the mask, while for gray-level images are filled with white.

As handwritten lines of text are frequently curved and rotated, causing, in extreme cases, a complete breakdown of recognition, line images are rectified by projecting every line segment of the text line image onto the horizontal axis. To achieve this, a piecewise affine transformation is estimated from points sampled at regular intervals along the baseline and mask boundary, along with their respective target coordinates after rotation. This transformation is subsequently applied to the cropped line image to produce an approximately straight and upright line, independent of curvature, rotation, and other distortions in the source image.

After every line is extracted from a given page, they are sorted according to the hash of their text[23]. Each image-label pair is further annotated with

[23] Due to limitations imposed by certain libraries, the line order of the page has not been preserved for certain manuscripts. This serves as a proxy for random shuffling while enabling a form of versioning.

six types of metadata associated with the manuscript, namely manuscript id or "shelfmark", century of production, language, script type[24], textual genre, verse, two types derived from the page layout with the region's and line's classes (e.g., Main text, Interlinear Line), and finally, metadata indicating which split the line belongs to (see Section 4.1).

3.2 Transcription Guidelines

Rationale Extrinsic variability, linked to the rendering of historical sources, and intrinsic variability, linked to the nature of the data, pose a twofold issue for automated transcription.

Firstly, in the landscape of philological studies, the focus on scholarly editions that represent the culmination of scholarly editing norms and practices has often overshadowed the importance of the transcription process, leading to intermediate imitative transcriptions relegated to drafts or even lost altogether. This prevailing notion of the edition as the ultimate goal has significantly influenced researchers' perspectives on transcription, blurring the line between the two, whether within or outside HTR settings. Ultimately, methods for transcribing or editing texts are shaped by local traditions and field-specific expectations, sometimes leading to isolated practices with contradictory outcomes, influenced by factors such as country, language, and project goal. Drawing upon the linguistic field of Old French for illustration, philologists within the Anglo-Saxon tradition, influenced by entrenched practices, may edit texts that are missing accents or apostrophes [6, Introduction, p. 5], particularly in Anglo-Norman contexts, resulting in forms such as *Dangleterre* and *labbe*, whereas literary edited texts from different traditions would typically yield *d'Angleterre* and *l'abbé*.

At the same time, the necessity to accommodate a wide array of writing styles, all while preserving a consistent approach (*cf.* Section 2) stands out as the most significant challenge. Several factors contribute to this such as: (i) the extensive temporal span, covering the 8^{th} to 16^{th} centuries, (ii) the diverse linguistic landscape characterized by dialectal variations and graphical systems with distinct special characters within each emerging or dominant language, and (iii) the multitude of document types, each adhering to its unique conventions.

To address these challenges, we present guidelines designed to produce straightforward yet precise transcriptions of historical sources in a computer-readable format. CATMuS adopts a strategic approach aimed at balancing information loss (in our case morphological variation of graphemes) with standardization in text reproduction. These guidelines, initially formulated for medieval French documents spanning from the 10^{th} to the 15^{th} centuries [39], and gradually incorporating more diverse documents, recognize that in practice, minor deviations

[24] Alongside CLaMM, we adhere to Derolez's typology [15], which is based on morphological aspects. However, it is worth noting that this typology is based on consensus, especially considering the substantial increase in script families and sub-families by the late 14^{th} century, as well as cursivity techniques and level of execution which also affect their visual characteristics.

from traditional transcription practices may be necessary to ensure and promote dataset coherence in a digital setting.

Character Representation and Word Segmentation Rather than opting for a wholly imitative transcription[25], we adopted a transcription approach that groups the forms of letters within the system of the Latin alphabet at the time. This approach, known as a graphemic approach [52], ensures a many-to-one mapping where allographs (or "character variants") are normalized, resulting in a distinctive representation for each letter:

- Graphic distinctions for the same character like ‹s› ("round ‹s›") and ‹ſ› ("long" ‹s›), or ‹ꝛ› ("round" ‹r›) and ‹r› ("short" ‹r›), are disregarded to mitigate potential cascading effects on other letters like a, d, e, etc.
- Ligatures between letters, such as ‹st› or ‹ct›, are treated as variations in form, leading to the independent transcription of their constituent characters.
- Any emphasis on the letter, regardless of its form, is transcribed as capital letters, including if the emphasis is only conveyed by size of the letter (and does not result in a different shape).

One of our most significant choices lies in what is called the "ramist" distinction between the pairs ‹u›/‹v›, and ‹i›/‹j›. We disregard this distinction, as these signs generally denoted variations in form (often tied to the character's placement in the word) rather than distinct phonetic realizations in most medieval sources. An illustrative example is the Roman numeral ".iij." (iii), where ‹j› is a prolonged ‹i› rather than the modern ‹j›. As a result, each ‹v› is transcribed as a ‹u›, and each ‹j› as an ‹i›.

Lastly, given that canonical word separation for all Latin and vernacular texts in Western Europe became standardized only by the end of the Middle Ages, when the word emerged as the primary unit of meaning, the representation of spaces between words in our dataset varies. It spans from *scriptio continua* (without spaces) to aerated text and deliberate use of lexical spaces. To alleviate potential discrepancies in imitative practices among transcribers, we opt for the semantic segmentation of words (lexical spaces). This decision, serving as the sole language-dependent choice in the project, has proven to yield beneficial results in sharing the guidelines among transcribers.

Abbreviations Adding to generalization issues as seen by Torres et al. [59], abbreviation resolution is often tied to local specificities and extra-textual information, which led us to identify abbreviation resolution as a natural language processing (NLP) task rather than an HTR one. For instance, the decision to develop the Tironian sign ‹⁊› into ‹et› or ‹e› depends on chronological and geographical factors and is considered an interpretation, as *e* is the Anglo-Norman variation of the Old French *et*. Additionally, abbreviation systems provide valuable insights

[25] A choice that, given the collaborative nature of our initiative, might have led to divergent interpretations of letterforms.

into the text, handwriting, production, or reception area, and text status (formal or working text). Preserving this information for resolution in a later stage of the textual acquisition pipeline can be highly beneficial.

The various signs fall into two main categories:

1. A large number are additions to base letters with diacritical marks, such as macrons (*e.g.* ‹ē› for *est*, *-em*, etc.), and superscript letters (‹qⁱ› for *qui*), and their meaning is often context-dependent. Diacritics are thus treated as separate characters using Unicode decomposed form (NFD).
2. Few are separate semantically distinct signs, such as ‹&› for *et/e*), or letters with strike-through marks, such as ‹ꝑ›. For these, we resorted to the Medieval Unicode Font Initiative (MUFI) [25].

Diacritics and special characters are treated similarly to letter allographs: for example, the macron and the horizontal tilde are both represented as a horizontal tilde, with the latter being chosen for ease of typing on most Western keyboards.

Punctuation, functional Signs, and Corrections Medieval documents feature intricate and varied punctuation systems, differing across periods, genres, and individual practices. Representing punctuation accurately requires an in-depth understanding of the manuscript. However, achieving a comprehensive representation of punctuation across multiple projects and transcribers is not feasible. To ensure consistency, four standardized signs[26] synthesize the complexity of punctuation systems:

- Single full stops are transcribed as ‹.›;
- Double signs are represented by ‹;›;
- Commas are transcribed as ‹,›,
- Question marks are transcribed as ‹?›, as soon as they appear in the Latin inventory in late middle ages.

Functional signs, when present in the source, serve to denote hierarchical segmentation of content (‹¶›), hyphenation (‹-›) for an end-of-line break of words, referral outside of the main text (‹⌗› - “dotted cross”), or insertion (‹˄›).

Finally, we have adopted a standardized approach to transcribe corrections noted in the sources, marked with expunction, strikethrough, or similar annotations. Adhering to the Leiden transcription conventions, the text identified by the scribal correction is enclosed within double brackets, namely ‹⟦›(U+27E6) and ‹⟧›(U+27E7).

3.3 Dataset Statistics

First of all, depending on the time of the digitization and the type of document scanned, often microfilms of the original documents (30 documents out of 203

[26] All punctuation marks are transcribed directly after the preceding sign without spaces.

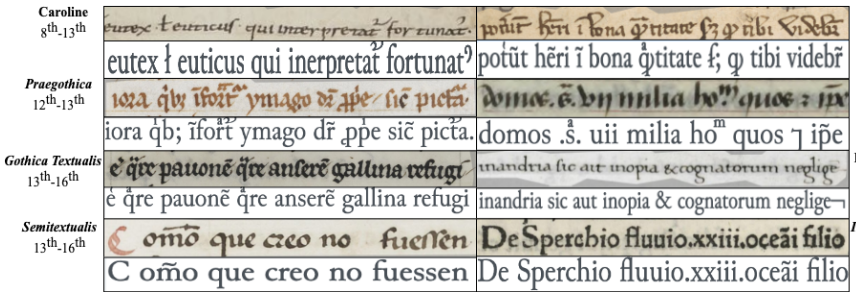

Fig. 3: Representative examples of the main script types lines, with CATMuS annotations.

are grayscale images), which are generally cheaper and less harmful to the original document, are used. Resolution quality within the dataset varies from one document to another.

The different metadata classes accompanying the dataset were chosen on the basis of historical relevance, accommodating historical analysis as well as facilitating testing of various architectural and model designs in one or combined research questions. On one hand, document information (script, shelfmark, century) provides graphical variation information. On the other hand, language-specific characteristics significantly influence the frequencies of certain characters[27]. Moreover, the literary genre and form, whether verse or prose, can influence various features such as line length, number of characters, punctuation usage, or script execution to some degree. Certain script types were also often favored depending on the genre of the text being copied (e.g., the *Textualis* script was typically used for copying the Bible). Lastly, the century of production serves as a generic proxy for language evolution, with potential impacts on lexical usage, conservation state, visual characteristics, and more.

In the following paragraphs, we provide comprehensive details about the dataset, including metrics concerning the number of lines, and character frequencies, potential biases inherent in the dataset, and the benchmarking possibilities it offers.

Word- and Character-level Analysis The distribution of the number of words and character per line, on average 7 words and 37 characters per line, seems to align with benchmarking datasets such as IAM and LAM, as well as exhibiting regularity as the length is bimodal-distributed (Figure 4). The full dataset comprises 220 classes of characters, including 25 combining superscript characters, 16 stricken-through characters, and 22 combining diacritics.

Script and Language Distribution Five projects [38,22,62,26] collectively contribute to around 64% of the dataset or 112,00 lines, resulting in an inherent

[27] For instance “k” and “w”, 90-95% of which is present only in text written in Middle Dutch and English.

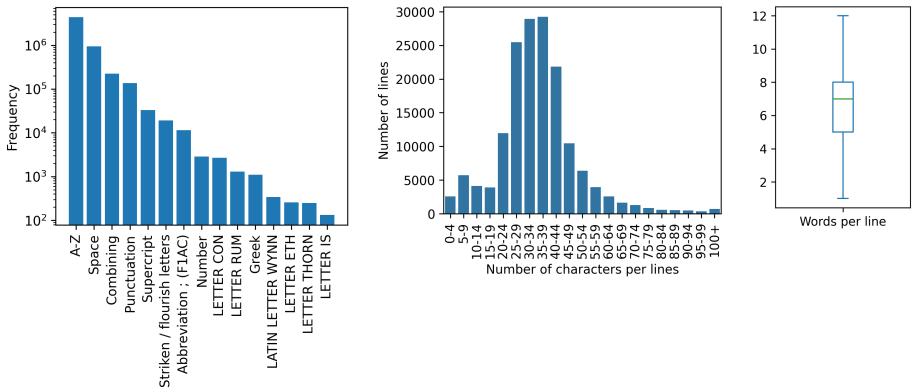

Fig. 4: Statistics on the frequency of character types (left), number of words (middle), and characters (right) per line.

overrepresentation of the scripts in which they are written, namely Caroline Minuscule, *Textualis*, and *Semitextualis* (cf. Figures 5). Similarly, the prolonged focus on Old French within these projects has led to its overrepresentation. Notably, the Herne project dataset alone accounts for 27% of the dataset, predominantly comprising Dutch from the 14$^{\text{th}}$ century in *Textualis*. Consequently, Old French, Latin, and Castilian together account for more than 50% of the dataset in *Textualis*. We integrated print into the dataset (approximately 12% of the dataset or 11,000 lines) because *incunabula* demonstrate an affiliation with manuscripts primarily, in terms of scripts, as existing handwriting shapes inspired the initial typographic characters and subsequently in terms of abbreviating and spacing practices.

Cross-features Benchmarking Capacities In terms of balanced representation across scripts, the Latin language stands out as the most suitable for benchmarking cross-script evaluation. Nearly all scripts are represented in Latin, except for *Semitextualis*. Castilian offers a strong representation, featuring six scripts out of ten possible for this language[28].

On the other hand, for cross-language benchmarking, *Textualis* emerges as the prime candidate, with all languages represented except for Old English. The same possibility is offered by the *Hybrida* and *Cursiva* scripts, though less prevalent[29], as they are shared among Castilian, Catalan, French, Italian, and Latin.

Lastly, since textual genres influence abbreviation practices and lexical choices, cross-genre evaluations could be conducted, particularly focusing on the Treatise

[28] Excluding the Caroline script, which historically ceased to be used as Gothic scripts came to the foreground in the 13$^{\text{th}}$ century.

[29] Especially for *Hybrida*, which offers a distinct Spanish variation.

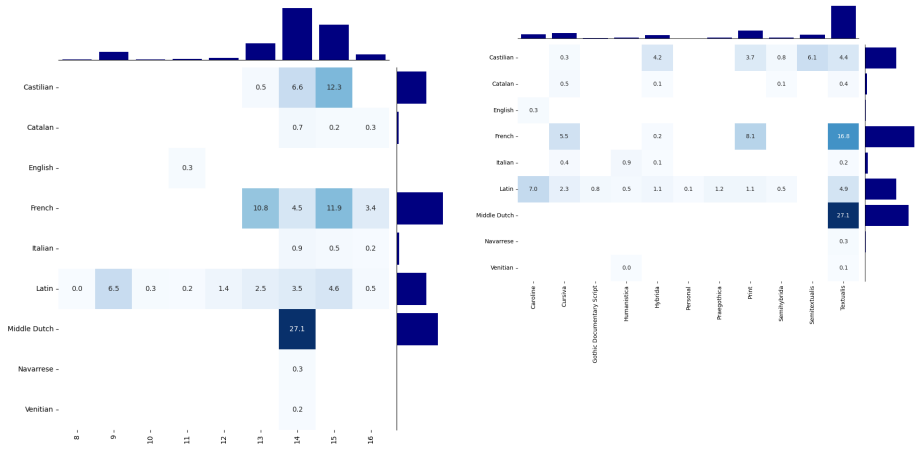

Fig. 5: Distribution of the dataset across languages, centuries, and scripts, in % of the total character count.

and Narratives genres. These genres encompass technical texts such as commentaries or medical treatises on one hand and historical or hagiographic on the other.

4 CATMuS Medieval for Benchmarking

4.1 Benchmark Splits

General Split The first split is designed so that every manuscript is present, with their lines divided on a 90% / 5% / 5% ratio. This allows for building models that have seen the full range of the data diversity, which is particularly important for languages less represented in the training data, such as Venitian (2 documents), Navarese (1), or Old English (1). This split is provided in the `gen_split` column of the parquet file.

Feature-based Split The second split is organized such that each manuscript is allocated to a single split, simplifying the assessment of model generalization. However, the data distribution across splits is uneven due to variations in the number of lines per manuscript. Data distribution was managed as follows:

1. Rare languages were distributed manually: English, having specific letters, was put in the train set only; Venitian has one manuscript in each of the train and test split; Navaresse, as a language of Spain, was put in the test set only, as it provides an interesting linguistic case (it is close to Castilian and Catalan);
2. Italian and Catalan (9 and 5 documents each) were split with one document in test, one in dev, and the rest in train. The choice of the document to put

Table 3: Statistics on the two different benchmarking splits.

Split	Set	Lines	Characters	Languages	Documents	Scripts
General	Train	146,885	5,178,330	9	203	10
	Validation	8,317	293,888	9	203	10
	Test	8,317	294,435	9	203	10
Feature-based	Train	138,900	4,892,639	8	176	10
	Validation	11,376	446,148	5	11	9
	Test	13,243	427,866	8	17	9

in dev and test was made based on the frequency of the script both in the larger corpus and the language class;
3. Each script was then picked from the dataset, to be present in train, dev, and test if possible.
4. Other documents were dispatched by language, using the number of characters to match roughly a 90% / 5% / 5% split, resulting in a 85% / 7% / 7% splits (see Table 3).

4.2 Baseline Scores for Handwritten Text Recognition

Setup In this section, we report an experimental analysis of the performance of two software behind the most common platform for HTR used in the humanities and in the GLAMs, *Kraken* (version 4.3.10) [32] and *Pylaia* (version 1.1.0) [45], on both splits. The performances are reported in terms of Character Error Rate (CER). This is the number of substitutions, deletions, and insertions that have to be applied to the predicted sentence to obtain the ground truth. In addition, we calculate the error rate associated with the space character, incorporating the count of inserted single dots, as they are frequently indicative of space-related errors based on our experience.

We trained *Kraken* and *Pylaia* with the default specifications or for the first and the recommended specifications for the later [36], in deterministic mode, with a batch size of 64, with three different seeds (21, 42, 84), and 5-epoch patience. Learning rate was set to $1e^{-4}$ for *Kraken* and $5e^{-4}$ for *Pylaia*. We trained on RTX8000, with 12 CPU cores for each training.

Results The results presented in Table 4 offer valuable baseline metrics for each split, shedding light on the trade-offs inherent in different model architectures and hyperparameters utilized by each toolkit. As anticipated, the "General" split exhibits lower CER, given the absence of out-of-domain documents, whereas the "Feature"-based split surpasses 10%. This higher score presents an intriguing challenge for developing more domain-specific models that consider factors such as script type and language.

In an effort to interpret the results, the most common sources of mistakes, independently from the split, are typically related to space and punctuation

Table 4: Baseline scores (average CER and standard deviation) with default configurations of *Kraken* and *Pylaia*. Each tool was used to train 3 models on each split. Validation CER is provided by each software, and Test CER was provided by Kraken's codebase. Time is based on the `real` of the unix `time` command. Space-related CER is computed on the test set with errors regarding space deletion, insertion, and single mark insertion.

			Character Error Rate (%)		
Split	Software	Training Time	Validation	Test	Space-related
General	Kraken	2112 min ± 163	5.7 ± 0.07	4.7 ± 0.06	1.0 ± 0.02
Feature	Kraken	1464 min ± 238	6.8 ± 0.16	13.1 ± 0.24	2.7 ± 0.06
General	Pylaia	308 min ± 047	9.1 ± 0.63	8.4 ± 0.73	1.8 ± 0.11
Feature	Pylaia	295 min ± 078	11.3 ± 0.24	21.2 ± 0.92	3.8 ± 0.06

and are directly related to the inconsistency of this practice in the documents (with ever-evolving languages). Another set of common errors is linked to the insertion and deletion of combining signs - this is an interesting mistake as it suggests that contextual noise can skew predictions for combining characters, usually those occupying the upper writing band of the line (*cf.* Figure 3).

5 Conclusion

In this paper, we introduced the CATMuS dataset tailored for line-level HTR of historical manuscripts in Latin scripts, comprising over 160,000 lines, which we are releasing on HuggingFace[30]. Alongside image-text pairs, it includes historical metadata, making it suitable not only for HTR research but also for tasks like script classification, century dating, and palaeographical analysis. Comprehensive analysis of the dataset, including quantitative and qualitative assessments of its characteristics and performance using two widely used architectures for HTR namely *Kraken* and *PyLaia*, underscore the challenges posed by the dataset's inherent variation. We believe these challenges contribute to making the dataset a valuable resource for advancing effective solutions in HTR for historical documents. As a future extension of this endeavor, enhancing the dataset mainly with more fine-grained labels on scribes and provenance could elevate its level of supervision and task-specific splits, broadening its suitability for additional tasks related to historical manuscript analysis. As a result of this joint collaboration, we are looking forward to improvements in the space of graphemic multilingual diachronic handwritten text recognition.

References

1. Alba, R., Rubin, G., Boschetti, F., Fischer, F., Clérice, T., Chagué, A.: HTRomance, Medieval Italian corpus of ground-truth for Handwritten Text Recog-

[30] See https://huggingface.co/datasets/CATMuS/medieval.

nition and Layout Segmentation [dataset] (2023). https://doi.org/10.5281/zenodo.8272751, https://github.com/HTRomance-Project/medieval-italian, v1.0.1
2. Aruta, D., Lenzi, M., Le Huërou, A., Possamaï, M., Pinche, A.: Liber [dataset] (2023), https://github.com/CIHAM-HTR/Liber, v0.0.5
3. Bhunia, A.K., Ghose, S., Kumar, A., Chowdhury, P.N., Sain, A., Song, Y.Z.: Metahtr: Towards writer-adaptive handwritten text recognition. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15825–15834 (2021). https://doi.org/10.1109/CVPR46437.2021.01557
4. Biay, S., Boby, V., Konstantinova, K., Cappe, Z.: TNAH-2021-DecameronFR [dataset] (2022). https://doi.org/10.5281/zenodo.6126376, https://github.com/PSL-Chartes-HTR-Students/TNAH-2021-DecameronFR, v1.0
5. Bordier, J., Gille Levenson, M., Brisville-Fertin, O., Clérice, T., Chagué, A.: HTRomance, Medieval Spain corpus of ground-truth for Handwritten Text Recognition and Layout Segmentation [dataset] (2023), https://github.com/HTRomance-Project/middle-ages-in-spain, v0.0.6
6. Breuil, E.: Méthodes et pratiques de l'édition critique des textes et documents modernes. No. 27 in Bibliothèque de littérature du XXe siècle, Classiques Garnier (Nov 2019). https://doi.org/10.15122/isbn.978-2-406-08639-0, https://classiques-garnier.com/methodes-et-pratiques-de-l-edition-critique-des-textes-et-documents-modernes.html
7. Buringh, E.: Medieval Manuscript Production in the Latin West. Brill (2010). https://doi.org/10.1163/9789047428640, http://booksandjournals.brillonline.com/content/books/9789047428640
8. Camps, J.B., Baumard, N., Langlais, P.C., Morin, O., Clérice, T., Norindr, J.: Make love or war? monitoring the thematic evolution of medieval french narratives. In: Computational Humanities Research (CHR 2023). pp. 734–756. CEUR-WS.org (2023)
9. Camps, J.B., Vidal-Gorène, C., Stutzmann, D., Vernet, M., Pinche, A.: Data diversity in handwritten text recognition. challenge or opportunity? In: Digital Humanities 2022. pp. 160–165 (2022)
10. Camps, J.B., Vidal-Gorène, C., Vernet, M.: Handling Heavily Abbreviated Manuscripts: HTR Engines vs Text Normalisation Approaches. In: Document Analysis and Recognition – ICDAR 2021 Workshops. pp. 306–316. Springer International Publishing (2021)
11. Cascianelli, S., Pippi, V., Maarand, M., Cornia, M., Baraldi, L., Kermorvant, C., Cucchiara, R.: The lam dataset: A novel benchmark for line-level handwritten text recognition. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 1506–1513. IEEE (2022)
12. Claustre, J., Smith, D., Torres Aguilar, S., Bretthauer, I., Brochard, P., Canteaut, O., Cottereau, E., Delivré, F., Denglos, M., Jolivet, V., Julerot, V., Kouamé, T., Lusset, E., Massoni, A., Nadiras, S., Perreaux, N., Regazzi, H., Treglia, M.: The e-NDP project : collaborative digital edition of the Chapter registers of Notre-Dame of Paris (1326-1504). Ground-truth for handwriting text recognition (HTR) on late medieval manuscripts. (Feb 2023). https://doi.org/10.5281/zenodo.7575693
13. Cloppet, F., Eglin, V., Stutzmann, D., Vincent, N., et al.: Icfhr2016 competition on the classification of medieval handwritings in latin script. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 590–595. IEEE (2016)
14. Clérice, T., Chagué, A., Vlachou-Efstathiou, M.: CREMMA Medii Aevi [dataset] (Oct 2023), https://github.com/HTR-United/CREMMA-Medieval-LAT, v0.1.2

15. Derolez, A.: The palaeography of Gothic manuscript books: From the twelfth to the early sixteenth century. Cambridge University Press (2003)
16. Fischer, A., Frinken, V., Fornés, A., Bunke, H.: Transcription alignment of latin manuscripts using hidden markov models. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. p. 29–36. HIP '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2037342.2037348
17. Fischer, A., Frinken, V., Fornés, A., Bunke, H.: Transcription alignment of latin manuscripts using hidden markov models. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. pp. 29–36 (2011)
18. Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., Stolz, M.: Automatic transcription of handwritten medieval documents. In: 2009 15th International Conference on Virtual Systems and Multimedia. pp. 137–142. IEEE (2009)
19. Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S., Litman, R.: Scrabblegan: Semi-supervised varying length handwritten text generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4324–4333 (2020)
20. Gabay, S., Pinche, A., Vlachou-Efstathiou, M., Christensen, K.: Données HTR imprimés du 16e siècle [dataset] (Aug 2022), https://github.com/Gallicorpora/HTR-imprime-16e-siecle, v0.0.43
21. Gauthier, L.: The Death of the Historical Corpus (Sep 2021), https://univ-paris8.hal.science/hal-03337341, working paper or preprint
22. Gille Levenson, M.: Towards a general open dataset and model for late medieval Castilian text recognition (HTR/OCR). Journal of Data Mining and Digital Humanities (2023). https://doi.org/10.46298/jdmdh.10416
23. Glaise, A., Clérice, T., Boschetti, F., Fischer, F., Chagué, A.: HTRomance, Medieval Latin corpus of ground-truth for Handwritten Text Recognition and Layout Segmentation [dataset] (2024), https://github.com/HTRomance-Project/medieval-latin, v0.0.6
24. Gueville, E., Wrisley, D.J.: Transcribing Medieval Manuscripts for Machine Learning (Sep 2023), https://shs.hal.science/halshs-03725166, working paper or preprint
25. Haugen, O.E.: MUFI Character Recommendation: Characters in the Official Unicode Standard and in the Private Use Area for Medieval Texts Written in the Latin Alphabet. Medieval Unicode Font Initiative, Bergen (2009), https://mufi.info/q.php?p=mufi
26. Haverals, W., Kestemont, M.: The middle dutch manuscripts surviving from the carthusian monastery of herne (14th century): Constructing an open dataset of digital transcriptions. In: Computational Humanities Research (CHR 2023). pp. 135–152. CEUR-WS.org (2023)
27. Hawk, B.W., Karaisl, A., White, N.: Modelling medieval hands: Practical ocr for caroline minuscule. Digital Quarterly Journal (January 2019), https://github.com/rescribe/carolineminuscule-groundtruth
28. He, S., Sammara, P., Burgers, J., Schomaker, L.: Towards style-based dating of historical documents. In: 2014 14th International Conference on Frontiers in Handwriting Recognition. pp. 265–270. IEEE (2014)
29. Kahle, P., Colutto, S., Hackl, G., Mühlberger, G.: Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 04, pp. 19–24 (Nov 2017). https://doi.org/10.1109/ICDAR.2017.307

30. Kang, L., Riba, P., Wang, Y., Rusinol, M., Fornés, A., Villegas, M.: Ganwriting: content-conditioned generation of styled handwritten word images. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16. pp. 273–289. Springer (2020)
31. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
32. Kiessling, B.: Kraken–an universal text recognizer for the humanities. In: Digital Humanities 2019 Conference Abstracts. Utrecht, The Netherlands (2019), https://dev.clariah.nl/files/dh2019/boa/0673.html
33. Kiessling, B.: A modular region and text line layout analysis system. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 313–318. IEEE (2020)
34. Kiessling, B., Tissot, R., Stokes, P., Ezra, D.S.B.: escriptorium: an open source platform for historical document analysis. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 19–19. IEEE (2019)
35. Leroy, N., Pinche, A., Camps, J.B., Clérice, T., Chagué, A.: HTRomance, Medieval French corpus of ground-truth for Handwritten Text Recognition and Layout Segmentation [dataset], https://github.com/HTRomance-Project/medieval-french, v0.0.7
36. Maarand, M., Beyer, Y., Kåsen, A., Fosseide, K.T., Kermorvant, C.: A comprehensive comparison of open-source libraries for handwritten text recognition in norwegian. In: Uchida, S., Barney, E., Eglin, V. (eds.) Document Analysis Systems. pp. 399–413. Springer International Publishing (2022)
37. Nikolaidou, K., Seuret, M., Mokayed, H., Liwicki, M.: A survey of historical document image datasets. International Journal on Document Analysis and Recognition (IJDAR) **25**(4), 305–338 (2022)
38. Pinche, A.: Edition nativement numérique du recueil hagiographique" Li Seint Confessor" de Wauchier de Denain d'après le manuscrit 412 de la Bibliothèque nationale de France. Ph.D. thesis, Thèse de doctorat, Lyon, Lyon (2021)
39. Pinche, A.: Guide de transcription pour les manuscrits du Xe au XVe siècle (Jun 2022), https://hal.archives-ouvertes.fr/hal-03697382, working paper or preprint
40. Pinche, A.: HTR Models and genericity for Medieval Manuscripts (Jul 2022), https://hal.science/hal-03736532, working paper or preprint
41. Pinche, A.: Cremma Medieval [dataset] (Oct 2023), https://github.com/HTR-United/cremma-medieval, v2.0.1
42. Pinche, A., Gabay, S., Leroy, N., Christensen, K.: Données HTR incunables du 15e siècle [dataset] (Oct 2023), https://github.com/Gallicorpora/HTR-incunable-15e-siecle, v0.0.28
43. Pinche, A., Gabay, S., Leroy, N., Christensen, K.: Données HTR manuscrits du 15e siècle [dataset] (Oct 2023), https://github.com/Gallicorpora/HTR-MSS-15e-Siecle, v0.0.37
44. Pinche, A., Pierreville, C.: Fabliaux [dataset] (Jun 2023), https://github.com/CIHAM-HTR/Fabliaux, v0.0.22
45. Puigcerver, J.: Are multidimensional recurrent layers really necessary for handwritten text recognition? In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 67–72. IEEE (2017)
46. Romero, V., Fornés, A., Serrano, N., Sánchez, J.A., Toselli, A.H., Frinken, V., Vidal, E., Lladós, J.: The esposalles database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognition **46**(6), 1658–1669 (2013)

47. Romero, V., Toselli, A.H., Vidal, E., Sánchez, J.A., Alonso, C., Marqués, L.: Modern vs Diplomatic Transcripts for Historical Handwritten Text Recognition. In: New Trends in Image Analysis and Processing – ICIAP 2019. pp. 103–114. Springer International Publishing (2019)
48. Sanchez, J.A., Romero, V., Toselli, A.H., Vidal, E.: Icfhr2016 competition on handwritten text recognition on the read dataset. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 630–635. IEEE (2016)
49. Serrano, N., Castro, F., Juan, A.: The rodrigo database. In: LREC. pp. 19–21 (2010)
50. Seuret, M., Limbach, S., Weichselbaumer, N., Maier, A., Christlein, V.: Dataset of pages from early printed books with multiple font groups. In: Proceedings of the 5th international workshop on historical document imaging and processing. pp. 1–6 (2019)
51. Siglidis, I., Gonthier, N., Gaubil, J., Monnier, T., Aubry, M.: The learnable typewriter: A generative approach to text line analysis (2023), https://arxiv.org/abs/2302.01660
52. Stutzmann, D.: Paléographie statistique pour décrire, identifier, dater... normaliser pour coopérer et aller plus loin ? In: Fischer, F., Fritze, C., Vogeler, G. (eds.) Kodikologie und Paläographie im digitalen Zeitalter 2 = Codicology and Palaeography in the Digital Age 2. pp. 247–277. BoD (2011), https://halshs.archives-ouvertes.fr/halshs-00596970, schriften des Instituts für Dokumentologie und Editorik
53. Stutzmann, D.: Les «manuscrits datés», base de données sur l'écriture (2017)
54. Stutzmann, D.: Words as graphic and linguistic structures. Word spacing in Psalm 101 Domine exaudi orationem meam (eleventh-fifteenth centuries). In: Les Mots au Moyen Âge – Words in the Middle Ages, pp. 21–59. No. 46 in Utrecht Studies in Medieval Literacy, Brepols, Turnhout (2020), 10.1484/M.USML-EB.5.120721
55. Stutzmann, D.: Fontenay dataset. original charters from fontenay before 1213 (2022)
56. Stutzmann, D., Aguilar, S.T., Chaffenet, P.: Home-alcar: Aligned and annotated cartularies (2021)
57. Stutzmann, D., Moufflet, J.F., Hamel, S.: la recherche en plein texte dans les sources manuscrites médiévales: enjeux et perspectives du projet himanis pour l'édition électronique. Médiévales pp. 67–96 (2017)
58. Torres Aguilar, S., Jolivet, V.: Handwritten Text Recognition for Documentary Medieval Manuscripts. Journal of Data Mining and Digital Humanities **Historical Documents and automatic text recognition** (Dec 2023). https://doi.org/10.46298/jdmdh.10484
59. Torres Aguilar, S.O., Jolivet, V.: Handwritten text recognition for documentary medieval manuscripts. Journal of Data Mining and Digital Humanities (2023)
60. Toselli, A., Romero, V., Villegas, M., Vidal, E., Sánchez, J.: Htr dataset icfhr 2016 (Feb 2018). https://doi.org/10.5281/zenodo.1164045
61. Vidal-Gorène, C., Camps, J.B., Clérice, T.: Synthetic lines from historical manuscripts: an experiment using gan and style transfer. In: International Conference on Image Analysis and Processing. pp. 477–488. Springer (2023)
62. Vlachou-Efstathiou, M.: Eutyches "de uerbo" glossed [dataset], https://github.com/malamatenia/Eutyches
63. Vlachou Efstathiou, M., Leroy, N., Maulu, M.: git-project-Boccace [dataset]. https://doi.org/10.5281/zenodo.6126613