Phrase-Level Segmentation on Medieval Corpora for Aligning Multilingual Texts

Lucence Ing*, Matthias Gille Levenson∝, Carolina Macedo◇
*ALMAnaCH (Inria), Paris, France
lucence.ing@inria.fr
∝UVSQ, DYPAC, Saint-Quentin-en-Yvelines, France - ENS de Lyon, CIHAM, Lyon, France
matthias.gille-levenson@ens-lyon.fr
◇École nationale des chartes / PSL, Paris, France - Biblissima+, Aubervilliers, France
carolina.macedo@chartes.psl.eu

Abstract

This paper presents an approach to multilingual alignment for medieval languages, focusing on the prior step of "phrase" segmentation. It outlines the challenges posed by historical data and describes different strategies for segmenting texts in multiple languages. It releases a gold-standard segmentation corpus based on various literary and historical works from the late Middle Ages in Europe. This corpus consists of texts in seven medieval languages (French, Castilian, Catalan, Portuguese, Latin, Italian, English). Several architectures are tested with both in-domain and out-of-domain evaluation sets.

Keywords: medieval texts, multilingual alignment, phrase segmentation, low-resource languages, digital philology, text transmission

1. Introduction

The recent development of large language models (LLMs) has been accompanied by increased interest in the study of multilingual textual traditions. From a Natural Language Processing (NLP) perspective, recent advances in multilingual sentence alignment—largely driven by deep learning and sentence embedding approaches (e.g., Liu and Zhu, 2023; Feng et al., 2022)—as well as in word-level alignment (Dou and Neubig, 2021), have opened new possibilities for the comparative study of textual traditions across multiple language versions and historical stages. These methods have already produced promising results in the field of stemmatology (Gille Levenson et al., 2024), where they support the analysis of textual relationships and the reconstruction of transmission histories.

At the same time, the increasing use of automatic handwriting recognition tools (Kiessling, 2019; Kiessling et al., 2019) and the emergence of generic medieval multilingual models (Clérice et al., 2024) make it possible now to consider the inclusion of corpora that are larger both in terms of the number of witnesses (i.e., extant manuscript copies or other textual attestations of a given work) included and in the overall volume of text processed, thereby enabling us to revisit complex historical traditions through computational methods.

Indeed, medieval Europe witnessed the emergence of literary works (i.e., abstract textual compositions transmitted through multiple manuscript witnesses) that were widely copied, translated and disseminated across regions and languages. Some of these works have survived the centuries, sometimes only as fragments, sometimes in more complete manuscripts. Faced with such diversity, how can we attempt to reconstruct the transmission history of a textual tradition? The *Lancelot en prose*, for instance, survives in more than 100 Old French witnesses. To what degree can medieval translations enrich our understanding of this tradition and contribute to the reconstruction of its transmission history? By contrast, in the case of the Post-Vulgate *Queste del Saint Graal*, whose Old French witnesses are rare and fragmentary, medieval translations play a central role, making it possible to reconstruct, to a large extent, an otherwise lost tradition.

In response to this challenge, within a broader multilingual alignment workflow tailored to medieval languages, we propose an approach that focuses on the initial phase: segmentation.

2. Segmentation as a Pre-Alignment Step

The alignment tool currently under development, *Aquilign*, is designed for medieval multilingual corpora and aims to explore relationships between texts from a philological perspective.The expected output is an alignment table in which similarities and differences between witnesses can be observed (see Table 1). The tool builds on *Bertalign* (Liu and Zhu, 2023), which assumes fixed upstream segmentation, i.e., that the segments to be aligned are predefined.

In this context, the segmentation task consti-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 936–946
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

French	Catalan	Castilian	English	Latin
Il conuient doncques	Coue donques	E por ende conuiene	ø	Oportet ergo
que la puissance en quoy **uertuz est mise** soit racionele.	que la potencia en la qual **ha esser la uirtut** sia racional.	que sea poderio razonable aquel en que **esta la uirtud**.	thanne **the vertue moot be** in the resonable vertue and potencial myght.	esse rationalem potentiam in qua **ponitur esse uirtus**

Table 1: Segment aligned across six versions (translations of Giles of Rome's *De Regimine Principum*, ca. 1280). French and Castilian follow the reading "*ponitur*", as found in the manuscripts, while Catalan and English introduce modal constructions.

tutes a crucial first step of the alignment pipeline: each witness is divided into linguistically coherent units that are then aligned. These pre-segmented texts are then encoded as sentence embeddings and compared across languages to identify high-probability correspondences. Medieval-aware segmentation ensures that alignment operates on syntactically and thus semantically meaningful units rather than arbitrary text segments. Our working assumption is that this improves alignment stability and limits incorrect matches between segments.

Since alignment relies on prior segmentation, a heterogeneous medieval training dataset for the segmenter becomes necessary (see Sec. 4).

3. From Modern Segmentation to Medieval Constraints

Sentence segmentation for contemporary languages is often considered a solved task, thanks to standardised orthography, stable punctuation, and the availability of large datasets. High-performing sentence splitters range from early unsupervised models such as Punkt (Kiss and Strunk, 2006) to recent punctuation-agnostic approaches (Minixhofer et al., 2023; Frohmann et al., 2024).

Few of these conditions hold for medieval texts: punctuation is unstable or partly editorial; graphic conventions vary across manuscript witnesses; corresponding passages often differ markedly in extent from one witness to another; boundary cues are weak—especially in diplomatic editions or transcriptions—and HTR/OCR (Handwritten Text Recognition / Optical Character Recognition) noise further compounds the problem. Consequently, segmentation is not only a computational task but also a philological one. Therefore specific methods adapted to the material features of medieval witnesses and to the particularities of historical multilingual traditions are required.

This challenge begins with the very notion of a "sentence" itself. In medieval witnesses, the modern concept of a "sentence" is not reliably encoded. The features discussed above—unstable or editorial punctuation and inconsistent capitalisation—do not consistently mark boundary onsets. To illustrate the difficulty of sentence boundary detection, consider the following example with two Old French transcriptions of the *same* passage from the *Lancelot en prose* (BnF fr. 111 and BnF fr. 751): periods may occur mid-clause, lowercase may follow a point, and initial capitals are unreliable.

> ***BnF fr. 111*** *: puis sen retourne tout contreual la riuiere pour sauoir sil pourroit trouuer ne pont ne gue Mais il ny en treuue point. Et quant il uoit quil ne puet passer si ne scet que faire car retourner ne uouldroit il pas. Adonc uoit yssir du chastel une damoisele*

> ***BnF fr. 751*** *: et puis sen torne tot contreual la riuere. Por sauoir se il i troueroit ne pont ne gue. mais il nen i truue point : si ne sot que faire car retorner ne wet il mie atant uoit issir une damoisele dou chastel*

This absence of reliable sentence boundary cues explains why rule-based splitters that rely on modern punctuation and capitalisation tend to underperform on medieval prose. These challenges have motivated the exploration of new methodologies adapted to historical languages states. For instance, (Bouma and Adesam, 2013) use a features-based token classifier to segment Old Swedish. Another line of work adapts contemporary NLP technologies to historical or manuscript-based data. (Rosensweig et al., 2025), for example, combine Masked Language Modelling (MLM) with Next-Token Prediction to improve segmentation performance on historical Hebrew texts. In this last project, the level of granularity is finer than the sentence, as it extends to "grammatically complete or semi-complete units".

The present work responds to the same underlying challenges, aiming to develop a segmentation approach suited to the specific characteristics of medieval texts. In particular, as in (Rosensweig et al., 2025), we aim for *phrase-level segmentation*— i.e., segmentation into grammatically coherent units smaller than full sentences — rather than a *sentence-level segmentation* approach. This finer granularity is of interest because it enables smaller-scale alignments between different witnesses, including cases where phrases are present in one witness but absent in another (omissions), and facilitates subsequent steps such as semantic similarity measurement, word-level alignment and witnesses clustering.

4. Dataset Description

The evaluation of segmentation methods relies on two complementary types of medieval data: an in-domain (ID) dataset and an out-of-domain (OOD) evaluation set consisting of distinct works, designed to assess generalisation beyond the training material.

4.1. In-domain Training Data

Motivated by the lack of reliable resources for medieval segmentation, we developed the *Multilingual Segmentation Dataset* (Ing et al., 2025)[1], a gold-standard resource for training and evaluating segmentation models on medieval texts. The corpus focuses exclusively on segmentation, providing manually validated segment boundaries without further normalisation. It covers seven medieval languages—Latin, French, Castilian, Portuguese, Catalan, Italian, and English—and spans the 13th to 15th century, with extensions into the mid-16th. The collection includes a wide range of prose genres, such as narrative, didactic, legal and theological works.

Training data were drawn from both diplomatic and critical editions, as well as specialised corpora (i.e., curated datasets containing transcribed or normalised historical texts; see dedicated bibliography, Section 13), thereby combining editorially normalised texts with more faithful, manuscript-oriented transcriptions. Because the dataset is designed to supply segmented units for training a segmenter, its linguistic coverage follows that of the multilingual traditions under investigation (e.g., the *Lancelot en prose* (Gille Levenson et al., 2024) and the *De regimine principum* (Ing et al., 2025)[2]).

Data acquisition, processing, and metadata curation Data collection proceeded on a language-by-language basis, reflecting the diverse conditions of access, preparation, and curation across sources. For traditions with readily usable resources—such as the *Base de Français Médiéval* (BFM) (Lavrentiev et al., 2011)—texts are already available in cleaned and structured formats (plain text and XML), requiring no additional preprocessing. Other traditions offer similarly rich materials but demand further normalisation and light text cleaning after download. This is notably the case for Portuguese sources such as CTA (Sobral, 2005) and CIPM (Xavier et al., 1993); for English, LAEME (Laing and Lass, 2008) and OTA (Oxford Text Archive, 1976–2021); for Italian and Latin, the *Biblioteca Italiana* (Baldi and Domenichelli, 2000); for Latin, ALIM (ALIM, 2003) and the *Latin Library* (Carey, 1997); and for Castilian, the Corpus of Hispanic Chivalric Romances (Corfis and Ancos, 2005) and CHARTA (Red CHARTA, 2015), among others.

Access is more restricted for certain corpora—such as CICA (Torruella and de las Heras, 1995), CORDE (Real Academia Española (RAE), 1998), and OVI (Larson et al., 1996)—which provide extensive and well-documented materials but only allow concordance-based querying, without direct access to downloadable texts. In such cases, textual data are either reconstructed through controlled extraction and manual compilation—as for the Castilian corpus CORDE—or supplemented with additional texts retrieved from alternative repositories and critical editions, particularly for Catalan.

Most sources are available in plain text or XML, with a few converted from HTML or PDF when no clean export is provided. Light, format-agnostic normalisation (whitespace, minimal structural clean-up) is applied while deliberately preserving historical orthography and punctuation.

To ensure methodological transparency and end-to-end traceability, we develop CorpusTemporis, a lightweight Streamlit-based web application that requires mandatory provenance fields (text, edition, editor, year, language, period, genre, edition type—critical/diplomatic/semi-diplomatic—source ID/URL). The application also provides a searchable *Texts* interface for editing and exporting curated records. Collected entries are stored locally in a consolidated CSV file, making data lineage auditable and facilitating reuse and replication across releases.

Following metadata consolidation, each language subset is compiled into a single plain-text files that serves as input to a custom script that randomly samples text chunks for manual segmentation. Files are then organised by language and include both raw and segmented versions, as well as monolingual and multilingual training splits.

Annotation methodology Following the automated sampling procedure described above, the extracted textual chunks are assembled into separate files organised by language, where manual segmentation is performed by annotators.

Segmentation does not rely on punctuation as a primary criterion, given its irregular and often editorial nature in historical texts. Nor is segmentation performed at the level of modern editorial "sentences", which are often not precise enough to capture the fine-grained correspondences required for alignment. Instead, segmentation targets smaller sub-sentential, clause-level units grounded in grammatical structure. Clause

[1] For detailed information, see the project repository.

[2] Project repositories: Lancelot en prose and De regimine principum

boundaries are established on the basis of linguistic evidence, primarily through syntactic and semantic criteria, reflecting the close correlation between syntactic organisation and units of meaning.

During the annotation process, each clause boundary is marked using a dedicated delimiter. The symbol £ serves this function, as it does not occur in the source texts and therefore avoids ambiguity during processing. Given the interpretive nature of this task, it cannot be reduced to a fixed set of rules or handled exclusively through regular expressions (see Sec. 5.1). The following examples illustrate typical segmentation outputs.

> *(Catalan) £él dix a la sua muler £que pugés en un caval, £per so car avien luyn as-anar, £per què ela tremolan £puyà al caval*[3]

> *(Italian) strenui, £e quasi del continuo guerreggiano co' Tartari precopensi ad essi vicini, £scorrendo nelle lor provincie £e predando i bestiami loro, £che poco altro in quei luochi £trovava che predare.*[4]

To ensure cross-linguistic consistency, segmentation follows a set of shared linguistic guidelines (see the project repository) defining clause-level boundaries. Within this framework, specific forms functioning as syntactic markers—such as relative pronouns, conjunctions, speech verbs, and adverbs, among others—serve as indicators of potential clause boundaries. Certain clause types likewise trigger delimiter insertion, including infinitival, gerund, or absolute ablative constructions. In such cases, morphological cues—typically verbal—signal the onset of a new segment. The guidelines further account for discourse-level phenomena such as direct speech and parenthetical insertions. The following examples illustrate conditional, gerundial, and parenthetical constructions that trigger delimiter insertion.

> *(Latin) £Si bonam conscientiam haberes, £non multum mortem timeres*[5]

> *(English) £havyng pety and compassyon of hys handwerke and hys creatur £turnyd helth into sekenesse.*[6]

> *(French) £Ne vos chaut, £fet ele, £vos le verroiz encore plus apertement £que vos nel veez ore*[7]

While the principles above define the contexts in which segmentation applies, certain cases are deliberately excluded to prevent over-fragmentation. Delimiters appearing at the end of a chunk without contextual continuation are ignored. When multiple potential delimiters occur in sequence, only the first is annotated in order to avoid redundant or semantically incoherent segmentation. Similarly, appositions and simple enumerations are not segmented, as they do not constitute independent propositional units. For example, *£Mas, da desaveença £que ouve antre mĩ e Moluca, o senhor de Calçom*[8], where the apposition *o senhor de Calçom* belongs to the same syntactic structure and therefore remains within a single segment. In this same example we also observe that the coordinated unit *mĩ e Moluca* is also not segmented. Coordination follows specific rules designed to balance linguistic precision and structural coherence. The adopted principle is that a conjunction linking two short elements—typically no more than two words—is not treated as a segmentation delimiter. When the coordinated elements are longer or internally structured, whether verbal or nonverbal, the conjunction is instead considered a delimiter. For instance, the example below illustrates both phenomena: one conjunction links two nouns and is not segmented, whereas another links two phrases and is treated as a segmentation boundary.

> *[...] (Castilian) £espandire sobre casado de Jacob* ***e*** *de David £****e*** *sobre los estageros de Jherusalem £spiritu de gracia e ruegos.*[9]

This distinction ensures that segmentation captures genuine clause boundaries while avoiding unnecessary fragmentation of coordinated phrases. Coordinating conjunctions are a frequent source of delimiter ambiguity across languages. As we have seen, certain tokens may function as segmentation boundaries or, depending on context, as simple connectors within a clause. This contextual ambiguity illustrates why segmentation cannot be addressed through regular expressions alone: the model must learn to discriminate between syntactically similar but functionally distinct cases. Section 6.1.3 details how the model handles such ambiguity.

[3] Footnote translations provide literal renderings of £-segmented units. Lit.: "he said to his wife that she should mount a horse, because they had far to go, so she, trembling, mounted the horse"

[4] Lit.: "tireless, and almost continually they wage war with the Precopensian Tatars, their neighbouring ones, raiding into their provinces and plundering their livestock, for little else in those places was found to plunder."

[5] Lit.: "If you had a good conscience, you would not fear death much".

[6] Lit.: "having pity and compassion for his handiwork and his creature, [he] turned health into sickness"

[7] Lit.: "Do not concern yourself, she said, you will see it again more clearly than you see it now".

[8] Lit.: "But because of the quarrel that arose between me and Moluca, the lord of Calçom".

[9] Lit.: "I will pour out upon the house of Jacob and of David, and upon the inhabitants of Jerusalem, a spirit of grace and supplications."

Language	Segments	Words	AWPS
Latin	14,056	67,935	4.8
French	12,168	78,391	6.4
Castilian	14,243	75,845	5.3
Portuguese	10,377	57,366	5.5
Catalan	7,983	51,631	6.5
Italian	7,764	50,778	6.5
English	6,970	36,155	5.2
Total	**73,561**	**418,101**	**5.68**

Table 2: Distribution of segments, words, and average words per segment (AWPS) across the seven languages in the dataset. Values reflect the relative size and structural granularity of the annotated material: higher AWPS indicates longer or more syntactically dense segments, whereas lower values correspond to finer clause-level segmentation.

Corpus statistics The annotated dataset, i.e., after manual segmentation, comprises approximately 73,000 segments, corresponding to about 418,000 words.[10]

Average segment length varies substantially across languages, reflecting the syntactic and editorial diversity of the sources—that is, variation in clause structure, genre, and rhetorical style, as well as in editorial conventions (e.g., normalisation and punctuation) that together influence both segment length and the detectability of boundary cues. In our dataset, the average number of tokens per segment (AWPS) is lowest in Latin and Portuguese (≈5.1–5.2) and highest in French and Italian (≈6.5), with Catalan and Old Castilian around ≈6.25 and Middle English at ≈5.9. As a compact overview, Table 2 summarises the distribution of segments, words, and average words per segment across the seven languages, highlighting cross-linguistic variation in corpus size and segment length.

Some variation in language representation is observed within the corpus. Although currently moderate, this imbalance highlights the importance of maintaining corpus balance, as even minor disparities can serve as early indicators of potential bias in model training. In its current state (release version 1.0), the corpus remains under active development, with ongoing efforts to expand the less represented subsets and move toward a more balanced cross-linguistic distribution.

4.2. Out-of-Domain Test Data

To assess generalisation across domains and languages, evaluations are conducted on HTR-derived transcripts and on languages not included in the training corpus.

[10]The corpus contains 534,478 tokens, including punctuation.

4.2.1. HTR Outputs

HTR has become an increasingly widespread method for text acquisition, and a growing number of projects work directly with such data, often with little or no manual correction or language normalisation. The out-of-domain (OOD) test set includes a subdataset made of HTR-derived transcripts with minimal correction and normalisation. Since HTR outputs is widely used today, we evaluate our method on data directly derived from automatic transcription. Evaluations on HTR outputs are therefore essential to assess how well the models generalise to "real-world" data. The evaluation covers the *Lancelot en prose* (Old French), Giles of Rome's *De regimine principum* (French, Latin, and Catalan versions), and the Castilian translation of Livy's *Ab Urbe Condita* by Pero López de Ayala (late 14th century). The selection of these texts is motivated by their availability and by their inclusion in projects directly associated with the present work. This evaluation setup challenges models trained on clean, edited text, to segment noisier, HTR-derived inputs. The total size of this subset is 6,281 words and 1,165 segments.

4.2.2. Closely Related Language

The second OOD evaluation targets a language closely related to one of the training languages, namely Occitan. Although genealogically and geographically close to Catalan, Occitan exhibits distinctive graphic and linguistic features, making it a suitable test case for assessing the model's ability to generalise beyond the training data. Performance on this language provides an indication of whether the system can extend to unseen but related varieties rather than merely reproducing patterns observed during training. The OOD corpus consists primarily of prose texts drawn mainly from the *Rialto* corpus (Girolamo, 2003), including the Occitan translation of the *Legenda aurea*, the pharmaco-medical treatise *Las vertutz de las herbas*, and other narrative or didactic works representative of the medieval Occitan tradition. The subset comprises 4,845 words and 859 segments.

5. Tested Methodologies

The first baseline is a regex-based engine. A second baseline is the Segment Any Text tool (Frohmann et al., 2024). These two methods are compared to three token classification architectures: a homemade Character-embedding Bi-LSTM and two Transformers-based networks, with two different base models: BERT[11] and Distil-

[11]Using the `BertForTokenClassification` method and the base model `google-bert/bert-base-multilingual-cased`.

BERT[12] which derives from the former.

Segmentation is only the first step of an overall multilingual alignment workflow. Therefore, the small size of the models is an important criterion, in order to provide a light and easy-to-use pipeline. This is the reason why no unilingual models were tested, as this would increase the size required to run the tool on a multilingual corpus tenfold. For supervised learning models, hyper-parameters optimisation is performed using Optuna toolkit (Akiba et al., 2019).

5.1. Rule-based Method

The basic approach for text segmentation is a rule-based approach, based on regular expressions, as it is done for example in pySBD (Sadvilkar and Neumann, 2020). Words that serve as segment boundaries are explicitly defined. This method is used as a base comparison for the others tested architectures. The efficiency of the rule-based approach is extended to its maximum, taking into account not only, for each language, simple patterns ("`\bque\b`", "`\bcomo\b`" ...), but also the graphical variation inherent in medieval languages ("`\b[Pp]er ?ci[òo] ?ch[eé]?\b`"), complex expressions ("`\b[tT]anto que\b`") and possible concatenation of simple patterns ("`\bcomo \b`" + "`\bque\b`").

5.2. Learning-based Methods

A token classification method Segmentation is framed as token-level boundary detection task. The tested classifiers assign each token a binary label indicating whether it begins a new segment (1) or continues the current one (0). A reconciliation of the sub-tokens produced by the tokenizers and the words (identified by punctuation and spaces) is performed.

Table 3 illustrates token-level boundary prediction: the model labels each token as either continuing the current segment (0) or starting a new one (1). This method has proven its worth in word segmentation tasks.[13]

This formulation decouples boundary detection from modern punctuation heuristics: the model learns language specific cues, producing segments that are suitable for downstream alignment, including on HTR-derived inputs.

[12] `DistilBertForTokenClassification` and `distilbert/distilbert-base-multilingual-cased`.

[13] For instance, Clérice (Clérice, 2020) explores an approach that encodes each character as an input unit, framing the task as word boundary classification. The objective is to process *scripta continua* texts, that is, texts written without spaces used to mark the different "words" of the sentence.

un	*chemin*	*viés*	*et*	*ancien*	***si***	*ne*	*demora*	*gueres*
0	0	0	0	0	**1**	0	0	0

Table 3: *si* is predicted as the beginning of a new segment (y=1); previous and following tokens are inside the segments (y=0).

The best model is chosen based on a weighted average of precision and recall.[14] A greater importance is given to recall, because the alignment phase offers the possibility of aligning m to n segments where $n >= 1$ and $m >= 1$. In other words, the next phase in the processing chain can compensate for over-segmentation (by merging segments), but it cannot for under-segmentation. Low false negative rate and, thus, a high recall rate, is preferred.

The chosen loss is the Cross-entropy loss implemented in Pytorch and Transformers libraries.

SaT We test Segment any Text (Frohmann et al., 2024), a tool to segment any text into sentences, without the help of punctuation. The SaT models are trained from Transformer models, initialized with the weights of XLM-RoBERTa (Conneau et al., 2020). They reach high F1-scores for the languages they can handle, as the best model reaches 93.1%. SaT (model `sat-12l`) is fine-tuned on our corpus, using the Low Rank Adapters (LoRA) feature provided by the software.

Character Embeddings and LSTM In medieval states of languages, the spelling is not fixed and numerous formal variations exist. HTR processing also adds graphic variation through the inevitable errors that models produce. To resolve the resulting out-of-vocabulary problems, we implement an architecture that produces Word Embeddings from Character Embeddings, using a series of convolution layers. This method is derived from (El Boukkouri et al., 2020). In this paper, the character to word embedding layer aims to replace wordpiece tokenisation in BERT models, to improve results on specific domains and vocabulary (namely, medical data).[15]

The chosen hyper-parameters for this architecture are: character embeddings dimension: 96, with a character dropout of 0.05. To the "final" word embeddings of dimension 768 is catted a lang metadata embeddings of dimension 8. The encoder is composed of 2 Bi-LSTM layers (hidden

[14] The chosen weights for precision and recall, based on several trials, are [1, 1.3] for BERT and LSTM-based models.

[15] Finetuning a model with the full architecture developped in (El Boukkouri et al., 2020) wasn't possible due to the absence of pre-trained multilingual CharacterBert model.

size 136) with multihead attention, and three fully connected layers for the classifier, with a hidden size of 256. The learning rate is 1.194e-3, and the chosen batch size 16.

BERT-based models The tested BERT-based (Devlin et al., 2019) models are the `ModelForTokenClassification`, that is a BERT encoder architecture with a linear classification head, with no modification on their architectures. For **BERT**, a learning rate of 3.8e-4, and a batch size of 64 are selected. For **DistilBERT** (Sanh et al., 2020-03-01) is tested as a lighter alternative to BERT. The chosen learning rate is 9.64e-05 and the batch size 32.

6. Results

6.1. In-domain

Two kinds of results are presented: first, global results of the different architectures, in Table 4; second, results per language, in Table 5. To ensure the comparability of results between different tokenisation strategies, all evaluations are conducted at word level.

6.1.1. Rule-based

Even with specific rules (cf. Sec. 5.1), the results for the rule-based approach are very poor: the F1-score reaches only 0.46 (Table 4). The classification of ambiguous words (e.g. *and*) as segment boundaries only, can explain a part of the poor results. For instance, in the phrase *the queen and the king*, *and* is not considered as a segment boundary in our annotations norms, as it belongs to the same semantic unit.

6.1.2. Learning-based

SaT SaT is evaluated with and without the LoRA adapter, using a segmentation threshold of 0.005. The default `sat-12l` model trained on contemporary data reaches a medium precision, with a score of 0.698 and a very low recall, with a score of 0.357: when the model identifies a token as a segment boundary, it is often correct, but it fails to identify most of the tokens that should be. The finetuned model reaches a higher recall score (0.677) with an important decay in precision (0.500).

The failure of correct identification can be explained by two reasons. First, the models address other challenges than ours, and they are trained to segment a text in full sentences, not phrases (coherent semantic and syntactically segments, shorter than a full sentence). Second, the format of SaT training data makes correspond one example to one segment only, which probably produces a lack of contextual information. In our own method, an example usually contains multiple segments. We have to split our examples in as many segments they contain, which results in a loss of context. This difference in data architecture should explain the differences between SaT and the results produced with the other architectures.

BERT / DistilBERT BERT based models show the highest results with in-domain data, with the highest recall. With a F1-score of 0.878 and 0.873, BERT and DistilBERT models produce close results with in-domain test data.

The per-language evaluation (Table 5) is performed only with the best BERT model. French and Portuguese have the highest scores, probably due to a very homogeneous way of annotating their data. English and Catalan, the languages that constitute the two smallest parts of the corpus, have logically the poorest F1 results.

Latin shows lower precision results, due to the difficulty of identifying recurrent words as delimiters in its text. This difference in precision could by explained by the fact that it is a language whose syntagms are mainly delimited by syntactic and morphological cues. For example, where the other languages use conjunctions and adverbs (*quand*, *when*, etc.) for circumstancial time phrases, Latin can use absolute ablative constructions that are not delimited by a specific word. For instance, in the following sentence: "*O fortunatam natam me consule Romam!*",[16] the delimiter for the beginning of the absolute ablative is *me*. Its use as a delimiter depends on the syntactic context.

6.1.3. Dealing with Ambiguity

The ability of the model to handle ambiguous tokens—i.e., tokens that may function as segment boundaries depending on context, as illustrated earlier in 4.1—is a key aspect in evaluating the quality of the meaning-based segmentation task. Table 6 shows the results of the best BERT model on *and*-equivalent coordinating conjunctions across languages.

The task is globally well achieved, showing the benefit of a semantic based-approach over a rule-based one. For the segment boundaries, greatest results are achieved on French, Castilian, Portuguese and English with a F1-Score of 0.9 or more. In Latin, for this classification, the model reaches a high precision (0.917) but a low recall (0.774), and the opposite, high recall and low precision for the segment content.

[16] "O happy Rome, born in my consulship!", translation by Jonathan G. F. Powell (Powell, 2015).

Model	Tokeniser & Embeddings	Precision	Recall	F1-score	Parameters
Regexp engine	ø	0.538	0.412	0.46	None
SaT	XLM-RoBERTa	0.698	0.357	0.472	2.77e+08
SaT + LoRA	XLM-RoBERTa	0.500	0.677	0.575	2.79e+08
BERT Model	BERT	**0.873**	**0.882**	**0.878**	1.77e+08
DistilBERT Model	DistilBERT	**0.872**	0.874	0.873	1.35e+08
LSTM Model	Character Embeddings	0.860	0.851	0.855	2.25e+07

Table 4: Comparison of models and tokenisation strategies. Results on test data (in-domain) for segment boundary labels. SaT + LoRA model is the model finetuned on our data.

Language	Precision	Recall	F1-score
French	0.916	0.945	0.93
Portuguese	0.9	0.861	0.88
Latin	0.848	0.888	0.868
Castilian	0.866	0.868	0.867
Italian	0.861	0.871	0.866
Catalan	0.861	0.854	0.858
English	0.845	0.865	0.855

Table 5: Per-language evaluation results of the best BERT model, for segment boundary labels, sorted by F1-score.

Some tokens know the same difficulty to be classified, as *&* in Catalan, *e* in Italian and *&* as segment content in Castilian. The presence of a high number of enumerations in the texts redacted in these languages may explain this phenomenon. Indeed, enumerations have to be split—following the annotation guidelines—, but don't correspond to a precise context that can be identified easily.

6.2. Out-of-domain

The results of the out-of-domain evaluation are shown in Table 7 and 8. These experiments are conducted on the three best architectures identified in the in-domain evaluation: BERT, DistilBERT and Character Embeddings LSTM. The LSTM model shows little ability to adapt to new data and to generalize, unlike BERT and DistilBERT. Both tables show clearly that BERT outperforms the other models.

HTR-derived With a recall of 0.804 and a precision of 0.867, the BERT model achieves the best results and demonstrates the best capacity for generalisation on noisy HTR data, three points ahead of DistilBERT and seven ahead of the LSTM-Character Embeddings model.

Similar language On the evaluated similar language, Occitan, close to Catalan (as described in 4.2.2), the BERT finetuned model still outperforms the other architectures, with a precision of 0.861 and a recall of 0.760.

7. Discussion and Further Work

This paper presents both a Multilingual Medieval Dataset and a new method for phrase-level segmentation of medieval texts. The models results seem to reach a limit in F1-score around 0.88 on in-domain data. This might be due to the size of the dataset, which is modest, as its production is particularly time-consuming. The task itself, which is not trivial and can lead to different interpretations depending on the annotators, could also explain those results. Nevertheless, the best model shows promising results on out-of-domain data, both in noisy HTR output, close to "real-world" data, and on languages similar to those in the corpus. Adding HTR-derived data to the training corpus is one of the next steps of this work.

The improvement of the segmentation model seem to have positive influence on the alignment phase: we evaluate on a fragment of the *Lancelot en prose* our alignment tool with a same Sentence Embedding model. The segmentation is performed with two different models: the current one, described in Table 4, and an older model trained previously on less data. Results can be observed in Table 9. The recall of the latest model improves by 8 points. This progress increases the number of alignment units by 14%. It decreases by 11% the mean number of segments per cell, and by 23% the standard deviation on this same value: the produced tables are longer and they contain smaller and more precise alignment cells.

Improving the accuracy of the phrase segmentation seems to improve the quality of the resulting alignment. This alignment could be further improved by fine-tuning the sentence-embedding model on an adapted corpus, which is currently an ongoing task.

8. Acknowledgements

We would like to express our warmest thanks to Marianne Reboul for her help updating SaT source code and implementing the evaluation strategy for this tool.

Language	Token	Segment content			Segment boundary		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
French	et	0.878	0.963	0.919	0.975	0.916	0.945
	e	1	1	1	1	1	1
Portuguese	e	0.793	0.860	0.825	0.926	0.886	0.906
Italian	e	0.805	0.785	0.795	0.845	0.861	0.853
	i	1	0.938	0.967	0.5	1	0.667
Castilian	e	0.865	0.8	0.831	0.932	0.960	0.944
	&	0.743	0.805	0.773	0.928	0.9	0.914
	y	0.762	0.941	0.824	0.976	0.891	0.932
Latin	et	0.762	0.912	0.830	0.917	0.774	0.840
English	and	0.760	0.9	0.824	0.950	0.870	0.909
Catalan	e	0.829	0.916	0.87	0.935	0.865	0.898
	&	0.5	0.667	0.571	0.778	0.636	0.7

Table 6: Results of the best BERT model on ambiguous tokens meaning *and*, per language. Some languages present several forms ; capitalised letters are not retained.

Model	Tokeniser & Embeddings	Precision	Recall	F1-score
BERT Model	BERT	**0.867**	**0.804**	**0.834**
DistilBERT Model	DistilBERT	0.845	0.772	0.807
LSTM Model	Character Embeddings	0.81	0.715	0.760

Table 7: Comparison of models and tokenisation strategies. Results on test data (out-of-domain, HTR data) for segment boundary labels.

Model	Tokeniser & Embeddings	Precision	Recall	F1-score
BERT Model	BERT	**0.861**	**0.760**	**0.808**
DistilBERT Model	DistilBERT	**0.862**	0.705	0.776
LSTM Model	Character Embeddings	0.768	0.622	0.687

Table 8: Comparison of models and tokenisation strategies. Results on test data (out-of-domain, Occitan sources) for segment boundary labels.

	Segmentation Model	Alignment units			
	Recall	Number	Mean	Median	Std
Former model	0.801	718	2.05	1	1.85
Latest model	0.882	820	1.82	1	1.51
Delta	+10.11%	+14.20%	-11.21%	0	-18.8%

Table 9: Evolution of the alignment results based on the increase in the segmentation model recall, with a stable precision. The “Number” column shows the number of rows in the alignment table. The mean, median and standard deviation are calculated on the number of segments per cell in the alignment table.

We thank the reviewers for their questions and remarks that helped improving the paper.

9. Ethical statement

The present work does not involve any particular ethical considerations, given that it focuses on medieval historical data. The pursuit of energy efficiency was one of the reasons for using methods that are relatively energy-efficient by current standards, and why it was decided not to test large language models other than BERT.

10. Data, models, and code availability

Our *Multilingual Segmentation Dataset* is available on Github: Multilingual Segmentation Dataset. The segmentation experiments can also be reproduced with the code available on Github : Aquilign. All the experiments and trainings were performed on the state `a0f1a64` of the data repository and `07b81964` of the code repository. The ‘main‘ branch contains the segmenter with the most efficient architecture. The best BERT model is publicly available on Huggingface: `https://huggingface.co/ProMeText/aquilign-multilingual-segmenter`.

11. Funding

This work has been partially funded by Biblissima+ consortium (id. ANR-21-ESRE-0005) and by the Inria “Défi”-type project COLaF. It benefited from the technical infrastructures of the Blaise Pascal Centre (CBP) at the École Normale Supérieure de Lyon (France) and of CLEPS infrastructure from Inria Paris.

12. Bibliographical References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631. ACM.

Gerlof Bouma and Yvonne Adesam. 2013. Experiments on sentence segmentation in Old Swedish editions. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013*, pages 11–26.

Thibault Clérice. 2020. Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin. 2020.

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 *Document Analysis and Recognition - ICDAR 2024*, pages 174–194. Springer Nature Switzerland.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Zi-Yi Dou and Graham Neubig. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128. Association for Computational Linguistics.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915. International Committee on Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891. Association for Computational Linguistics.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941, Miami, Florida, USA. Association for Computational Linguistics.

Matthias Gille Levenson, Lucence Ing, and Jean-Baptiste Camps. 2024. Textual Transmission without Borders: Multiple Multilingual Alignment and Stemmatology of the "Lancelot en prose" (Medieval French, Castilian, Italian). In *Proceedings of the Computational Humanities Research Conference 2024*, volume 3834 of *CEUR Workshop Proceedings*, pages 65–92. CEUR.

Lucence Ing, Matthias Gille Levenson, and Carolina Macedo. 2025. Premiers jalons de collation multilingue du *De Regimine Principum* latin et vernaculaire. Presentation at the Congrès international de linguistique et de philologie romanes. 30 juin–5 juillet 2025.

B. Kiessling, R. Tissot, P. Stokes, and D. 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.

Benjamin Kiessling. 2019. Kraken - an Universal Text Recognizer for the Humanities.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Lei Liu and Min Zhu. 2023. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. 38(2):621–634.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. Association for Computational Linguistics.

Jonathan G. F. Powell. 2015. Tullius Cicero, Marcus, poems, the famous orator Cicero. In *Oxford Research Encyclopedia of Classics*.

Elisha Rosensweig, Benjamin Resnick, Hillel Gershuni, Joshua Guedalia, Nachum Dershowitz, and Avi Shmidman. 2025. Automatic text segmentation of ancient and historic hebrew. In *Proceedings of the Second Ancient Language Processing Workshop (ALP 2025) associated with NAACL*, pages 1–11. Association for Computational Linguistics.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020-03-01. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.

13. Language Resource References

ALIM. 2003. *ALIM: Archivio della Latinità Italiana del Medioevo*.

Andrea Baldi and Mario Domenichelli. 2000. *Biblioteca Italiana*. Università di Roma La Sapienza.

William L. Carey. 1997. *The Latin Library*.

Ivy Corfis and Pablo Ancos. 2005. *CHCR: Corpus of Hispanic Chivalric Romances*.

Costanzo Di Girolamo. 2003. *Rialto: Repertorio informatizzato dell'antica letteratura trobadorica e occitana*.

Lucence Ing and Matthias Gille Levenson and Carolina Macedo. 2025. *Multilingual Segmentation Dataset for Historical Prose (13th–16th c.)*. Zenodo.

Margaret Laing and Roger Lass. 2008. *Linguistic Atlas of Early Middle English (LAEME)*. University of Edinburgh.

Pär Larson and Elena Artale and Diego Dotto. 1996. *OVI: Corpus OVI dell'italiano antico*.

Alexei Lavrentiev and Sophie Prévost and Bernard Colombat. 2011. *Base de Français Médiéval (BFM)*. ENS de Lyon, UMR 5191 ICAR, 1.0.

Oxford Text Archive. 1976–2021. *Oxford Text Archive (OTA)*.

Real Academia Española (RAE). 1998. *Corpus Diacrónico del Español*.

Red CHARTA. 2015. *CHARTA: Corpus Hispánico y Americano en la Red: Textos Antiguos*.

Cristina Sobral. 2005. *Corpus de Textos Antigos (CTA)*. Centro de Linguística da Universidade de Lisboa (CLUL).

Joan Torruella and Isabel de las Heras. 1995. *Corpus Informatitzat del Català Antic (CICA)*. Universitat Autònoma de Barcelona (UAB).

Maria Francisca Xavier and Maria de Lourdes Crispim and others. 1993. *Corpus Informatizado do Português Medieval (CIPM)*. Centro de Linguística da Universidade de Lisboa (CLUL).