

# Textual Transmission without Borders: Multiple Multilingual Alignment and Stemmatology of the “Lancelot en prose” (Medieval French, Castilian, Italian)

Matthias Gille Levenson<sup>1,2,3,\*</sup>, Lucence Ing<sup>1,3,\*</sup> and Jean-Baptiste Camps<sup>1,3</sup>

<sup>1</sup>Centre Jean Mabillon, École nationale des chartes, Paris Sciences & Lettres, France

<sup>2</sup>CIHAM, UMR 5648, École Normale Supérieure de Lyon, France

<sup>3</sup>ÉquipEx Biblissima+

## Abstract

This study focuses on the problem of multilingual medieval text alignment, which presents specific challenges, due to the absence of modern punctuation in the texts and the non-standard forms of medieval languages. In order to perform the alignment of several witnesses from the multilingual tradition of the prose *Lancelot*, we first develop an automatic text segmenter based on BERT and then align the produced segments using Bertalign. This alignment is then used to produce stemmatological hypotheses, using phylogenetic methods. The aligned sequences are clustered independently by two human annotators and a clustering algorithm (DBScan), and the resulting variant tables submitted to maximum parsimony analysis, in order to produce trees. The trees are then compared and discussed in light of philological knowledge. Results tend to show that automatically clustered sequences can provide results comparable to those of human annotation.

## Keywords

Multilingual alignment, Text segmentation, Medieval Arthurian literature, Stemmatology

## 1. Introduction

The production and transmission of written texts during Antiquity and the Middle Ages involved a process of manual copying. During this process, the text was progressively transformed by errors and innovations as it circulated, each copy introducing successive modifications. Since the 19th century, philologists have taken to study the transmission of texts, based on innovations, using the genealogical tree (*stemma codicum*) as a metaphor to visually represent this transmission process. Yet, the study of textual transmission is still often limited to the copies produced in a given language (e.g. Medieval French) and do not necessarily encompass all the translations, in other medieval languages, that were part of the history of a given work. This is due in part to stark difficulties in aligning and analysing multilingual traditions, but

---


CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

\*These authors contributed equally.

✉ matthias.gille-levenson@ens-lyon.fr (M. Gille Levenson); lucence.ing@chartes.psl.eu (L. Ing);

jean-baptiste.camps@chartes.psl.eu (J. Camps)

ORCID 0000-0001-9488-5986 (M. Gille Levenson); 0000-0002-8742-3000 (L. Ing); 0000-0003-0385-7037 (J. Camps)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

poses the risk of giving us a very limited and partial view of medieval cultures, constrained by linguistic “frontiers” that were actually quite blurry and porous at the time.

### 1.1. The challenge of multilingual traditions

Multilingual collation of texts has attracted a few contributions, mostly focused on encoding or displaying editions of widely circulated texts [24, 2, 20, 26]. Yet, due to technical and modelling challenges, most critical editions are limited to a single language. National academic traditions, and the relative rarity of comparative studies in, for instance, current Romance Philology might be a factor, since research is often focused on editing either the source text or a specific language version. This philological reality can also be explained by material issues, namely the great difficulty of taking into account a large number of witnesses in the editorial process, the elaboration of a stemma, and the production of editions. Already true for large unilingual traditions alone, this is even more acute for multilingual traditions of widely circulated works, since adding distinct language versions only increases the complexity of the task.

The use of computer tools can help to overcome some of these limitations and pave the way for global studies of multilingual traditions, and even for the production of multilingual editions of textual traditions in the longer term.

The interest of this approach lies in the history of the text, in a global romanistic approach: the aim is to consider the textual tradition as a whole, and to contribute to the progress of knowledge on the text in general. Connecting local traditions will be profitable in order to produce global knowledge on the reception of these multiply translated texts, while possibly improving local knowledge on a specific tradition. Multilingual collation could, for example, help clarify the history of the text in cases where one translation is the direct archetype of another.

This paper focuses on computational multilingual alignment and collation (subsection 2.2), at “sentence” level and on *similar* fragments, on the classification of aligned sequences into distinct variants (subsection 2.3), and on their subsequent stemmatological analysis (section 3).<sup>1</sup> To perform the alignment, due to the nature of the textual data we use, a first step of automatic segmentation of the text is necessary (subsection 2.1). The paper concentrates on the study of the multilingual tradition of the *Lancelot en prose*, in order to understand the history of the text and its translations. We take this text as a case study, which enables us to present a new methodology that can help the study of various multilingual medieval traditions.

### 1.2. The multilingual tradition of the *Lancelot*

The study is based on a few witnesses of a complex medieval tradition, that of the *Lancelot en prose*, an anonymous text composed in the first third of the 13<sup>th</sup> century which enjoyed great success throughout the medieval period, with at least 126 manuscript witnesses, followed by many printed editions from 1488 onwards [13, 4]. This great success is also evidenced by the translations it has undergone (into Castilian, Catalan, Italian, High German, and Dutch). The

---

<sup>1</sup>This article sets apart the identification of macroscopic displacement of large and medium-sized fragments (at paragraph level, for example) that can occur in complex traditions (see subsection 1.2).

alignment of texts in a multilingual framework allows, through the recording of variants, to establish the tradition of these translations.

In this paper, we focused solely on the Romance tradition, in particular the French source, and the translations in Castilian and Italian (with the exclusion of the Catalan translation, of which only small fragments remain). For both of these translations, only one complete witness is preserved: in a manuscript produced in Florence in the last quarter of the 14<sup>th</sup> century, for the Italian *Lancelotto* (Firenze, Biblioteca della Fondazione Ezio Franceschini 1); in a 16<sup>th</sup> century manuscript, copied from a 1414 exemplar according to its colophon, for the Castilian *Lanzarote* (Madrid, Biblioteca Nacional de España, 9611). Both texts have been edited [5, 9].

Given the extremely large number of witnesses of the French *Lancelot*, we have selected a sample of five witnesses. They have been chosen on the basis of their supposed relationships with the translations, according to existing philological knowledge: Paris, BnF, fr. 111 (15<sup>th</sup> c.) is supposed to be close to the *Lancelotto* [6], while Paris, BnF, fr. 751 (13<sup>th</sup> c., 2<sup>nd</sup> half), or more precisely the family of which it is a part, would be close to the *Lanzarote* [9]. In addition, due to their easy availability and well known versions, we added several reference points, in the form of the edition by Sommer, based on ms. London, BL, Add. 10293 (beg. 14<sup>th</sup> c.) [22, 23] and Micha, based on mss Cambridge, Corpus Christi College, 45 (13<sup>2/2</sup>) for volume 2 and Oxford, Bodleian Library, Rawlinson D. 899 (14<sup>th</sup> c.) for volume 4 [17, 18]. Finally, as a representative of the late French tradition, whose place in the genealogy remains to be elucidated, we included the *incunabula* edition of 1488 (from exemplar Paris, BnF, RES-Y2-46 and RES-Y2-47; Rouen, Jean le Bourgeois et Paris, Jean Dupré, 1488). For a list of the witnesses, see Appendix A.

The *Lancelot* is a very long prose text, and therefore it can be highly unstable from one witness to another. The witnesses containing the translations are not spared from this instability, and they are also fragmentary. *Lancelotto* is especially so, as it presents only three non-consecutive episodes of the text. To enable alignment, the first step was therefore to identify corresponding passages from one witness to another (Appendix B), and to retain only what could be compared. This is why the studied text segments have identifiers: ii-48, ii-61, and iv-75, corresponding to the sections of the text as they appear in Micha edition.<sup>2</sup>

### 1.3. State of the art on sentence alignment

While multilingual alignment is a fairly active field in NLP, relatively little work has been done on the production and use of alignment methods for philological, stemmatological and ecdotic purposes, in the field of heritage text with pre-orthographic languages.

Birnbaum and Eckhoff [3] are considering the creation of a multilingual alignment tool, based on parts of speech only, which works in cases of literal translations, such as the Old Church Slavonic and an original Greek version of the *Codex Supralensis*. The work of Meinecke, Wrisley, and Jänicke [16] on French epic literature, explores the possibilities offered by the use

---

<sup>2</sup>The first part of the identifier corresponds to the volume, the second to the number of the first segment according to Micha [17, 18] Due to significant textual variation and the current capabilities of the aligner, we have subdivided the longest sections, ii-61 and iv-75, into two and six parts respectively, so that none of the segments exceed 1,000 tokens. Despite this choice, the translations exhibit a significant number of omissions that are difficult to detect automatically. For example, *Lanzarote* contains the episode of the sparrowhawk (segment ii-61-2), but in an extremely shortened version, only a few sentences (Appendix C).

of semantic representations of words through embeddings, from a unilingual perspective only. However, nothing is said about the definition and identification of “sentences” in the source texts, as the unit chosen to compare the versions seems to be the verse. Yet, embedding-based similarity calculations seem to be a promising venue for text alignment [27].

Recently, Liu and Zhu [15] have published a tool called Bertalign, a two-steps algorithm that makes use of sentence-transformers and multilingual sentence embeddings, based on the LaBSE model (“Language agnostic BERT Sentence Embeddings”) [1]. In particular, it allows to align 1 to  $n$  and  $n$  to 1 texts fragments. Bertalign works with segments fixed upstream and is designed for modern language states.<sup>3</sup>

## 2. Methodology

In this paper, we use the fixed fragments method designed by Liu and Zhu and we make use of Bertalign to perform the global alignments, while at the same time proposing a specific segmentation approach prior to the alignment.<sup>4</sup> We describe our processing chain: segmentation, pseudo-sentences alignment, classification into distinct variants and stemmatological analysis.

### 2.1. Segmentation

The alignment task requires fixed segments, as the tool we rely on, Bertalign, is designed for contemporary languages for which sentence segmentation is not a problem (the period “.” is enough to split the corpus, as modern translations tend to reproduce the same divisions sentence by sentence). It can be an issue for medieval languages, where the notion of a modern sentence, which begins with a capital letter and ends with a period, doesn’t really exist. Punctuation is highly variable depending on the copyist, even if a global and comparative study has yet to be carried out. Let’s take the following two sentences as examples:

**BnF fr. 111** : *ains sem part si tost quil leut commandee a dieu et cheuauche en telle maniere iucqua tierce tant quil uient a lissue de la fourest. et il tourne a destre vers ung chemin uieil et ancien.*

**BnF fr. 751** : *Ains sen part si tost comme il lot comandee a dieu. et cheuauche en tel maniere. tant quil uient a lissue de la forest et il torne a destre uers le chemin uiez et ancien*

**Translation** (for both passages): [*But he goes away as soon as he commanded her to God and rides [until 9 a.m.], until he comes to the entrance of the forest, and he turns right towards an old pathway.*]

The punctuation of these two transcriptions is original: it shows that it is not a reliable marker for syntactic segmentation, as it can vary greatly depending on the source, and this

---

<sup>3</sup>A more recent preprint picks up on this work and compares Bertalign to a new architecture [14], but we haven’t had time to look at how the algorithm works and how good it performs on our data.

<sup>4</sup>The word-alignment phase, because it requires its own methodology, will not be dealt with in this article: we’re only interested in the “macroalignment” phase, at the sentence or syntagm level.

variability increases during the transcription and editing of the text. Moreover, original punctuation is still very little taken into account by editors, who re-punctuate the text and can make highly variable choices. It is therefore necessary to divide the texts into equivalent segments that can then be compared and aligned. To do this, we have decided to base our approach on sentence syntax. The assumption here is that there is a certain degree of correspondence between syntactic and semantic units, and that the syntactic tokenisation should be an efficient way to approach the different semantic units of the text. Let's take an example (Table 1): in the second line, "si" is translated by "e" in Castilian, but the syntactic units are preserved, allowing the alignment of segments that are formally quite different but share the same overall meaning.

**Table 1**

A correspondence between syntactic and semantic segments in fragment ii-48, on two following segments. The first one corresponds to the end of the preceding example. Each cell represents an alignment segment

| Micha   | Sommer                                       | Lanzarote   | Lancellotto  |
|---|--|---|--|
| et il torne a destre<br>vers un chemin viés<br>et ancien; | et torne a destre en<br>vn petit chemin viez | e dexo el camino<br>e tomo a mano<br>derecha vn camino<br>pequeño   e biejo<br>lleno de yerba | Ed e torna a destra<br>inverso uno camino<br>vecchio e antico, |
| si ne demora gueres                                       | Si ne demora gaires                          | e non andubo mucho  | si non dimora guari  |

### 2.1.1. Segmentation methods

To segment the texts, we try, compare and evaluate two methods: the first uses regular expressions, the second one uses AutoModelForTokenClassification BERT models.

**Using regular expressions** The first method we tried was based on the identification of function words, which required a language-by-language analysis of syntactic delimiters. Regular expressions have been developed to facilitate this identification and manage the high degree of graphical variation. Here are a few examples of such markers in French:

"[Tt]ant que", "[Aa]nsi", "[sS]i est car", "[sS]i est que", "[sS]i",  
"[Ee]t", "mais", "qu[ie] si", "[kc]ar", "[Qq]u?ant", "quar", "[Qq]u'",  
"[Qq]u[ei]", "[Qq]uel", "[Dd]on[ct]", "[pP]uis"

In Spanish:

"[Pp]ues que", "[pP]or?que", "[pP]er", "[Qq]ue", "a que", "si",  
"[Dd]o", "[Ee]", "comme quier que", "mas", "ass?i como", "como",  
"[Aa]n?ss?i", "para", "aquel que"

And in Italian:

"[mM]a si", "[Ss]i", "tanto che", "che", "e", "dove", "ch'?"

This regular expression method is limited, however, as it doesn't allow for fine segmentation in the case of repeated markers (you have to think of all possible combinations), which leads to over-segmentation of the text, as we'll see below.

**Creating an ad hoc segmenter** Another problem with rule-based segmentation is that it relies solely on the identification of character strings. Besides the fact that this leads to a proliferation of rules for languages with significant graphical variation (*ainsi, insi, einsi* for *thus*, for instance), it also identifies tokens that do not serve to delimit coherent syntactic units. Indeed, for example, the coordinating conjunction *et* (*and*) is a good delimiter of clauses and sentences, but it also serves to coordinate elements within the same syntactic units.

Let's have a look at the following two sentences:

- “*dont n'est a il mie merveille, se vos en estes couroucie **et** je vos pri que vos me diez que ce est*” [so it is not a surprise if you are angry **and** I beg you to tell me for what]: the conjunction *et* (*and*) coordinates two propositions and can be used as a delimiter;
- “*si me parti ier matin dolante **et** couroucie*” [I left yesterday morning sad **and** angry]: the conjunction coordinates two adjectives; it must not be used as a delimiter.

This is why the use of language models for identifying delimiters seemed relevant. Indeed, these models allow for capturing the semantics of different tokens, and thus differentiating the uses of the same words. It was decided to use AutoModelForTokenClassification BERT models [28], with the aim of identifying tokens that can serve as delimiters of these textual units. These models are usually trained on contemporary languages, but, as we will see, the large amount of data used to train them allows for good generalization and good results on our ancient language states.<sup>5</sup>

To produce specific training data for our corpus, tokens identified as delimiters were labeled with the value 1, all other tokens with the value 0, based on human evaluation and a specific set of grammatical rules.<sup>6</sup>

### 2.1.2. Segmenter train corpus

The models were trained on three languages: French, Castilian and Italian, in their medieval forms. The medieval Castilian corpus comprises around 8,000 tokenised lines (about 36,200 words), half of which are various texts from the 13<sup>th</sup> to 15<sup>th</sup> centuries, retrieved from the CORDE linguistic database [19]. The other half of the corpus consists of fragments taken from the *Lanzarote*, which are not included in the corpus dealt with in this paper. The French and Italian models were trained solely on data from the *Lancelot*, which is debatable from a model generalization perspective but was motivated by the desire to quickly achieve effective models for our texts. They were trained on 12,700 and 28,000 words (1,000 and 1,500 lines),

---

<sup>5</sup>The models which were used are: for the Castilian, the BETO model [8]; for the French and for the Italian, the models from the MDZ Digital Library [21].

<sup>6</sup>We retain a certain number of specific tokens, such as coordinating conjunctions between two clauses, relative pronouns, certain adverbs that begin sentences or clauses, etc., trying to segment appropriately without over-segmenting, for example, when two delimiter tokens appear consecutively (in the clause “*et qui trop bien se defren-doient*”, we only retain the conjunction *et*).

respectively. The difference in the number of words required to produce convincing models is due to the size of the initially used BERT models (the French BERT model is much larger as it was trained on a much larger dataset, making it more efficient for the given task).

### 2.1.3. Results

The results of the best segmentation models are presented in Table 2, Table 3 and Table 4, one for each language, according to classic measures of accuracy, precision, recall, and F1 score<sup>7</sup>. Each table displays the results of a model established using regex and a BERT-based model. The results show a substantial improvement in the text segmentation task, thanks to the use of a BERT-based model compared to the use of simple regexes.

The best model for each language is chosen on the basis of a weighted average between precision and recall, on a test corpus that has never been used for training. Given the whole workflow, recall is the important metric. It's more important to identify as many true delimiters as possible than to be precise in identifying delimiters. This is because the alignment phase that follows enables the alignment units to be merged, thereby compensating for false positives. Conversely, false negatives (failure to identify a delimiter) will not be compensated for later on, and will lead to an alignment of poorer quality. In view of this, it was decided to produce a weighted average between recall and precision for the selection of the best model, and to assign a weight of 2 to recall, and 1 to precision. Accuracy is high because it computes the total number of correct label assignments, both for labels 0 (non-delimiter tokens) and 1 (delimiter tokens), across the texts.

It is the second label, "Delimiter", corresponding to the results of identifying tokens labeled 1, that is relevant. Between the F1 scores of the regex and BERT-based models, we observe a significant improvement, with an increase of over 0.20 for French, 0.24 for Italian, and 0.23 for Spanish. Our models thus successfully segment the text according to a syntactic (and semantic) logic.<sup>8</sup>

**Table 2**

French model results of the segmentation models

|           | Regexp |           | BERT-based   |              |
|-----------|--------|-----------|--------------|--------------|
|           | None   | Delimiter | None         | Delimiter    |
| Accuracy  | 0.954  |           | <b>0.984</b> |              |
| Precision | 0.940  | 0.812     | <b>0.990</b> | <b>0.874</b> |
| Recall    | 0.978  | 0.611     | <b>0.978</b> | <b>0.941</b> |
| F1-score  | 0.959  | 0.670     | <b>0.984</b> | <b>0.906</b> |

We can see a short example of the segmentation produced by both methods:

**BERT-based:** *et bien paroit a ses iauz [SEP] qu'ele avoit rouges et anflez [SEP] qu'ele eust ploré*

<sup>7</sup>The evaluation was performed on test sets consisting of 4,300 words in Spanish, 1,340 words in French and 2,500 words in Italian.

<sup>8</sup>Due to the small size of the training corpus, it was decided to produce adhoc models for each of the languages in the corpus. A multilingual model with language metadata injection experiments will be produced and evaluated in an article dedicated to segmentation. These results provide a baseline for future in-depth studies.

**Table 3**  
Italian model results

|           | Regexp |           | BERT-based   |              |
|-----------|--------|-----------|--------------|--------------|
|           | 0.961  |           | <b>0.983</b> |              |
| Accuracy  | None   | Delimiter | None         | Delimiter    |
| Precision | 0.937  | 0.711     | <b>0.981</b> | <b>0.827</b> |
| Recall    | 0.971  | 0.523     | <b>0.976</b> | <b>0.866</b> |
| F1-score  | 0.954  | 0.602     | <b>0.978</b> | <b>0.846</b> |

**Table 4**  
Castilian model results

|           | Regexp |           | BERT-based   |              |
|-----------|--------|-----------|--------------|--------------|
|           | 0.951  |           | <b>0.981</b> |              |
| Accuracy  | None   | Delimiter | None         | Delimiter    |
| Precision | 0.942  | 0.688     | <b>0.981</b> | <b>0.869</b> |
| Recall    | 0.962  | 0.584     | <b>0.982</b> | <b>0.863</b> |
| F1-score  | 0.952  | 0.632     | <b>0.981</b> | <b>0.866</b> |

**Regex:** *et bien paroit a ses iauz qu'ele avoit rouges* [SEP] *et anflez qu'ele eust ploré*

The regex segmentation cut the sentence once, in the middle of a phrase (“*rouges et anflez*” that is “*red and swollen*”), whereas the BERT-based segmentation correctly segments the sentence twice, in accordance with the grammatical structure, based on relative pronoun and subordinative conjunction (*qu'*). The method chosen also has an influence on segment size, which is not shown in the evaluation above, and which has an impact on the quality of the alignments produced, in favor of the BERT-based method, as we'll see below.

## 2.2. Alignment

Once the BERT-based models are trained and selected, they are integrated into the alignment workflow, producing text segmentation based on the recognition of the good delimiters, before the actual alignment phase. As stated above, Bertalign is used to perform the alignments<sup>9</sup>. The alignment is carried out on the basis of a main/pivot witness chosen in advance on the basis of philological knowledge, that is, Micha [17]. This pivot witness is compared with each of the others, and the alignments are merged. To create the merged alignments table we make use of a graph method to connect each pair of aligned segment and create the final alignment unit, while conserving the 1 to many and many to one alignments. Considering each aligned pair as connected nodes in a graph, it is possible to build up complete alignment units by connecting all nodes together thanks to the pivot fragment<sup>10</sup> (figures 1 and 2).

<sup>9</sup>The parameters we used are the following: `max-align=3, window=5, skip=-.2, margin=True, len_penalty=True`.

<sup>10</sup>As with all methods using a pivot witness, alignment units in which the pivot omits text (omission or innovation of other branches) must be handled subsequently. Indeed, if the base witness is omitted, the link between the other witnesses is unknown. They must therefore be re-injected, which would require to design a second round of alignment that would change the base witness on the omitted parts. This work has yet to be produced.



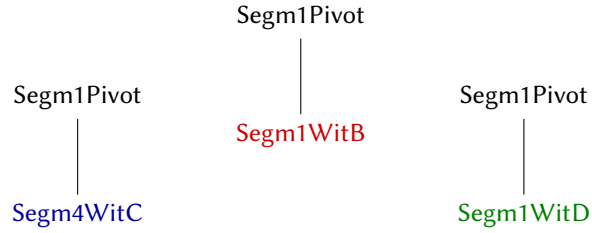


Figure 1: Getting pairs of aligned fragments

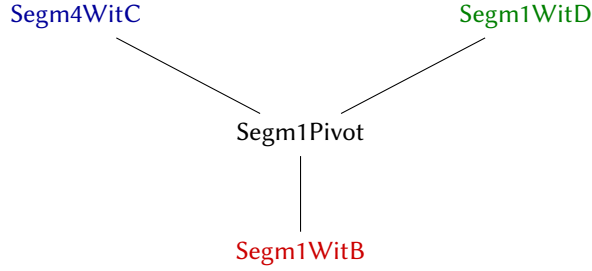


Figure 2: Merging the pairs into a single alignment unit by connecting all the nodes

### 2.2.1. Alignment evaluation

To assess the impact of the chosen segmentation method on alignment, we conducted an evaluation of the alignment produced by each of the two methods (regex and BERT-based segmentation).

**Description of alignment results** The evaluation is based on correcting the alignments obtained (correcting the alignment table of indices). Given that segmentation varies between different segmenters, the correction process must be repeated for results from both BERT-based segmentation and regex segmentation. Table 5 shows some alignment results with the fragment iv-75-1 (see also Appendix D, Table 11).

These results present the same alignment error in *Lancelloto* and BnF fr. 111, but due to two different causes: in the case of the *Lancelloto*, the error can be attributed to the segmentation phase, where the segmenter saw a single unit and did not separate between “*consiglio*” and “*chi*”. In the case of BnF fr. 111, segmentation was done properly, but the error was produced in the alignment phase, where the two units were regrouped.

**Evaluation** Alignments are evaluated on the basis of the Alignment Error Rate (AER) [25, 2.6, p. 21]:

$$AER = 1 - \frac{2 \times |L \cap L_{gold}|}{|L| + |L_{gold}|}$$

---

but it concerns a small number of alignment units only.

**Table 5**

Example of alignment automatically produced, with regex segmentation on the segment iv-75-1 (small extract). The meaning of the passage is “...that I would remedy it as I can”. We can see an alignment error with witness BnF fr. 111.

|  |   |   |  |  |   |   |  |  |
|--|---|---|--|--|---|---|--|--|
| que<br>metroie<br>volentiers<br>tout le bon<br>conseil | j'i<br>e<br>buon<br>siglio<br><b>potrei</b> | lancellotto<br>tutto il<br>con-<br>ch'i | fr333<br>que ie i<br>metroie<br>uolenters<br>tot le bon<br>conseil | inc<br>que ie y met-<br>troie tout le<br>meilleur con-<br>seil | fr111<br>car ie y<br>mectroye<br>uolentiers<br>tout le con-<br>seil  <b>que ie<br/>pourroye</b> | fr751<br>et uolentiers<br>i metroie<br>tout le bon<br>conseil | sommer<br>que ie y me-<br>terioie volen-<br>tiers tout le<br>bon conseil | lanzarote<br>que os<br>porne todo<br>el mejor<br>consejo |
| que je por-<br>roie                                    |   |   | que ie por-<br>roie  | que ie pour-<br>roie   |   | que ie i por-<br>roie mestre<br>y porroie<br>mettre           | que iou<br>y porroie   | que yo<br>pudiere<br>mettre                              |

Where  $L$  and  $L_{\text{gold}}$  consist of predicted and hand-corrected links. The evaluation is conducted on indices by comparing automatically generated indices with corrected ones. As the alignment is produced for each witness compared to the pivot witness, an AER is produced for each pair, and Table 6 presents the average of those in order to have an overall value for each evaluated segment of text.

We distinguish between two types of results: results that consider incorrect alignments due to segmentation issues, and those that do not. The evaluation is performed on different parts of the text, each with varying numbers of alignment units.

**Table 6**

Alignment Error Rate with LaBSE, excluding or including segments that were not properly segmented. The numbers of evaluated segments are in the “Nb of segm.” columns. Numbers in bold are detailed *infra* (improvement of the results with BERT-based segmentation for ii-48 and ii-61-1 sections but not for the iv-75-1 one).

| fragment | Regex-segm.    |      | Nb of segm. | Bert-segm.     |             | Nb of segm. |
|----------|----------------|------|-------------|----------------|-------------|-------------|
|          | well segmented | all  |             | well segmented | all         |             |
| ii-48    | 0.13           | 0.16 | 71          | <b>0.11</b>    | <b>0.15</b> | 112         |
| ii-61-1  | 0.30           | 0.38 | 136         | <b>0.26</b>    | <b>0.32</b> | 152         |
| ii-61-2  |                |      |             | 0.15           | 0.18        | 101         |
| iv-75-1  | <b>0.24</b>    | 0.33 | 120         | <b>0.27</b>    | 0.32        | 169         |
| iv-75-2  |                |      |             | 0.25           | 0.26        | 102         |

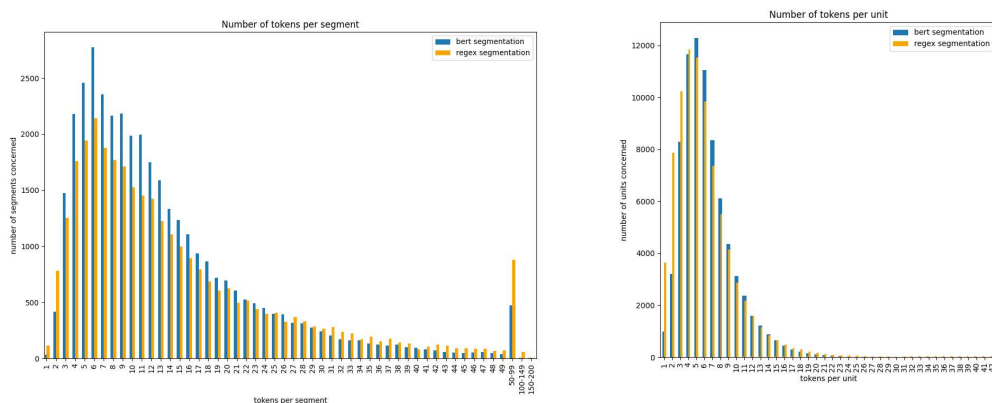
We observe a systematic improvement between the results that consider erroneous segmentation and those that do not, both for regex segmentation and BERT-based segmentation. Several sections of text were evaluated using both methods, revealing a slight overall improvement in results with BERT-based segmentation.

It is important to note that the quality of alignment varies depending on the textual proximity among versions in each witness. For example, there is a gap (0.15) between the results for sections ii-48 and ii-61-1, which can be attributed to varying lacunas and omissions observed in several witnesses for the latter.<sup>11</sup>

<sup>11</sup>For example, *Lancellotto* doesn't present the text for the beginning of Agloval episode, that leads to a poor alignment, because around 20 alignment units are not present in this witness.

Moreover, the poor results observed for the iv-75-1 section (better results are observed for regex segmentation than for BERT-based segmentation) can be attributed to numerous instances of missing correspondence among several witnesses, reflecting divergent traditions with frequent omissions. In this context, regex segmentation, generating fewer segments (669 compared to 797 in this case), yields better alignment. This is because these segments consolidate more units, potentially accommodating omissions that might affect other segments. In contrast, BERT-based segmentation, which offers finer granularity, tends to make more errors.

There is indeed a notable difference in alignment results due to the size of the produced data: regex segmentation generates less segments, i.e., aligned units, as illustrated in Figure 3.



**Figure 3:** Number of tokens per segment (aligned units, left) and unit (small units automatically produced by the segmenter, right). BERT-based segmentation produces units with a more stable number of tokens and thus yields more semantically coherent aligned segments.

BERT-based segmentation results in a significantly higher number of segments containing 3 to 14 tokens (even until 21 tokens), which correlates with the higher total number of segments obtained with this method. Conversely, regex segmentation produces more segments with a higher number of tokens (more than 30), as well as segments with one or two token(s). These segments are either too large to facilitate accurate alignment or too small to maintain semantic coherence. Thus, alignment tends to be of better quality with the BERT-based method for stemmatologic purposes.

Correlatively, the size of the units (the units produced by the segmentation) within the segments changes, as we can see from the figure 3. With the regex segmentation, there is a high number of very small units containing only one, two, or three tokens. In contrast, there are more units containing 5 to 8 tokens with BERT-based segmentation. As the units in regex segmentation are very short, they are grouped easily into segments. If this facilitates better results for the highly variable parts of text we described above, this approach results in a generally poor alignment, whereas BERT-based segmentation produces fewer segments with more semantically consistent units. If the results from the AER are not very promising, we can assume that the BERT-based segmentation produces a more coherent segmentation that positively impacts the alignment task. It is important to note that AER is a good metric for evaluating the quality of an alignment, but that it is not sufficient. Indeed, a text can be split into only three

parts and have a low AER, if the splits are correct, whereas this kind of result would not allow for any meaningful exploitation of the data. On the contrary, the evaluation of the distribution of units and segments lengths shows that BERT-based segmentation provides more coherent elements to align.

## 2.3. Classification in variants

### 2.3.1. Method

Most stemmatological and phylogenetic algorithms needs features (i.e. variants) that are arbitrarily coded (usually, using letters or numbers), e.g. variant 1, variant 2, etc. [7] In our case, this necessitates, for each alignment unit, to cluster the readings from every witnesses into sets of relevant clusters (Table 7). The number of these clusters can vary from 1 (all witnesses have the same variant) to a maximum number equal to the total number of witnesses (each witness bears a different variant).

**Table 7**

Table showing an aligned portion of text for each witness, and its numerically coded classification in variants by both annotators (here in agreement) and a clustering algorithm. The main opposition here is between “*heard*” (*oy, udito...*) and “*done*” (*fait*), and an omission. An OCR mistake (“*or*” for “*oi*”) in part prevents a correct clustering by the algorithm.

| Witness     | Aligned text        | Ann. 1 | Ann. 2 | DBScan |
|-------------|---------------------|--------|--------|--------|
| Micha       | qu’il avoit or      | 1      | 1      | 4      |
| Sommer      | quil auoit oy       | 1      | 1      | 1      |
| fr751       | quil auoit fait:    | 2      | 2      | 2      |
| fr111       | quil auoit ouy      | 1      | 1      | 1      |
| inc         |                     | 0      | 0      | 0      |
| lanzarote   |                     | 0      | 0      | 0      |
| lancellotto | ch’elli aveva udito | 1      | 1      | 3      |

As such, this task is an unsupervised clustering task, where the number of desired clusters cannot be known in advance, which excludes most clustering methods (such as, for instance, k-means or k-medoids). For this reason, we choose an unsupervised density-based clustering method, DBScan (Density-based spatial clustering of applications with noise [11]), that we apply to the cosine distances between each witness, for each alignment unit, in the embedding.

DBScan tries to find dense clusters of points, i.e., points with a given minimum number of neighbours (*MinPts*) situated at a given maximum distance ( $\epsilon$ ). If DBscan does not necessitate to set the number of clusters, it requires the setting of the two parameters *MinPts* and  $\epsilon$ . These parameters can only be chosen contextually, in relation to the number of individuals (witnesses) and the distances between them. For the *MinPts*, we set it at 1, given the small number of witnesses. For  $\epsilon$ , we follow a methodology called DMDBScan, that tries to identify relevant values, by first computing the minimum distances to the *MinPts* nearest points, and then plotting them by ascending order to identify values for at which there are sudden sharp changes in the minimum distances, possibly revealing different densities [10]. Following this methodology, we set  $\epsilon$  at 0.2 (Appendix F).

**Table 8**  
Mean Adjusted Rand Index

|             | Ann. 1 | Ann. 2 | DBScan |
|-------------|--------|--------|--------|
| Annotator 1 | 1      |        |        |
| Annotator 2 | 0.70   | 1      |        |
| DBscan      | 0.04   | 0.04   | 1      |

DBScan is applied to the cosine distances between the readings of the witnesses in the embedding. Yet, due to the multilingualism of the corpus, and the important spelling variation characteristic of medieval language, some distances are artificially increased not for semantic reasons but because of formal bias. To correct for this bias, a weighting is applied to all distances prior to performing clustering, according to:

$$\text{weightedDist}(a_i, b_i) = \frac{\text{dist}(a_i, b_i)}{\frac{1}{2}(\mu(\text{dist}(a, n)) + \mu(\text{dist}(b, n)))}$$

where  $a$  and  $b$  are two witnesses, and  $i$  a given alignment unit, and where  $\mu(\text{dist}(a, n))$  is the arithmetic mean of the distance between  $a$  and all other witnesses for all segmentation units.

### 2.3.2. Results and evaluation

In order to evaluate the results of clustering, we compare the clustering performed by DBScan to that of the two human annotators (the first two authors of this paper), and compute the mean Adjusted Rand Index (ARI), whose value is contained between -1 and +1, where -1 denotes a complete opposition, 1 a complete agreement, and 0 a case where both clustering appear to have been independently randomly labelled.

We compare the mean ARI between the two human annotators and the DBScan result. The results are presented in Table 8. The score is very low for the DBscan results, especially compared to the correspondance between the human annotators. Yet, it does not necessarily mean that the clustering results does not yield significant genealogical information to be processed by the stemmatological algorithms, as will be seen in the next section.

## 3. Results of the stemmatological analysis

The alignment tables we obtained through the BERT-based segmentation and the alignment phase enable philological analysis. It was performed in two distinct ways: first, a traditional stemmatic analysis, based on human expertise, was performed. Then, the variants resulting from the human annotations and the DBScan clustering were subjected to stemmatological algorithms, and the results were compared.

### 3.1. Human-identified witness groups

We decided first to evaluate the links between the witnesses, based on the alignment tables and human close reading expertise (ours, and that of previous philologists having worked on

this topic). Each one of the witnesses we chose presents particular readings (subsection E.1). However, we can distinguish two main groups within the tradition. On one hand, the witnesses Micha, BnF fr. 751, BnF fr. 111 and *Lancellotto*, and on the other hand, Sommer, the incunable, and *Lanzarote*. They can be determined on the basis of omissions or distinct reading groups (subsection E.2). The oppositions between the two groups are quite frequent and stable.

Nevertheless, *Lanzarote* has common readings with the first group and sometimes oscillates between the readings of the two groups (subsection E.3). In fact, the text in the *Lanzarote* presents a quite original version, especially due to the shortened passages it contains (subsection 1.2). It also often offers specific readings and innovations, particularly when the translation presents cases of direct speech instead of basic narration (subsection E.4).

The text of the *Lancellotto* doesn't show such originality. On the contrary, it is very close to the French text, using many gallicisms [5, p. 43-44]. It also stays very close to its own group. It shares particular common readings with BnF fr. 111, as well as several omissions. However, sometimes, *Lancellotto* knows a common reading with Micha against BnF fr. 111 (subsection E.5). The *Lancellotto* also shows some innovations that prove that the very model of the translation is one not part of our selection (subsection E.6).

We can assume that the witnesses we studied belong to two main groups, and that the witnesses that interest us, the *Lanzarote* and the *Lancellotto*, each fall into one of these groups. However, the specific variants and innovations found in each witness do not allow the identification of a specific model or the determination of a more precise affiliation.

The identification of groups confirms the filiation between BnF fr. 111 and *Lancellotto* [5], but does not confirm the specific link between BnF fr. 751 and *Lanzarote*. It is important to note that BnF fr. 751 is really unstable. Indeed, if it shares most of its readings with the first group, in the variants we studied, it also has its own unique readings.

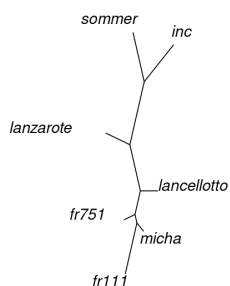
### 3.2. Results of the stemmatological algorithm

The second approach applies a well known phylogenetic algorithms to the table of variants produced by the two human annotators, and the one produced by the DBScan clustering. The chosen algorithm is the Branch and bound algorithm, whose goal is to determine minimal evolutionary trees by the criterion of maximum parsimony [12]. This algorithm can produce one or several trees, if different configurations achieve the same maximum parsimony value. The trees resulting from the analyses are presented in Figure 4.

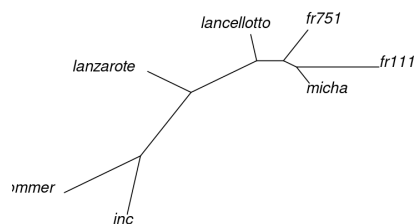
The trees resulting from the tables produced by the human annotators present, unsurprisingly, strong similarities with the result of the human close reading expertise on the genealogy. In particular, they all display the groupings {*Lanzarote*, Sommer, Incunabula} versus {*Lancellotto*, Micha, fr. 751 and fr. 111}. The second and third alternative trees from Annot. 2 also show the grouping of *Lancellotto* and fr. 111, already observed by the editor of the *Lancellotto* [6], but only the third tree from Annot. 2 shows the stronger similarity between fr. 751 and *Lanzarote*, that was hypothesised by its editor [9].

In comparison, the results based on the DBScan do not show fully the opposition {*Lanzarote*, Sommer, Incunabula} versus {*Lancellotto*, Micha, fr. 751 and fr. 111}, because Sommer switches families. Yet, it provides results more consistent with the analyses of the editors of the *Lanzarote*, that is positioned close to the fr. 751 [9], while *Lancellotto* remains relatively close to

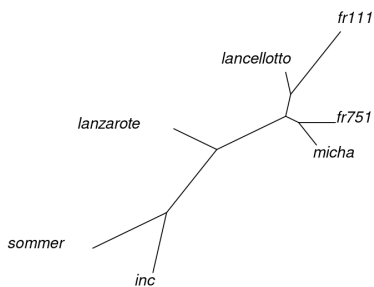
Unrooted single tree - Annotator 1



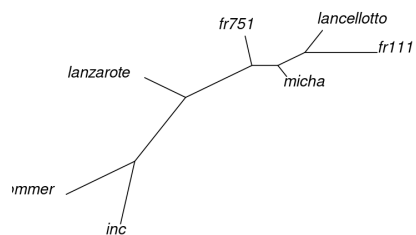
Unrooted alternative tree - Annotator 2



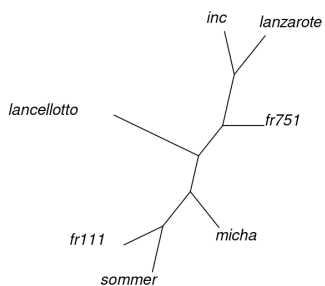
Unrooted alternative tree - Annotator 2



Unrooted alternative tree - Annotator 2



Unrooted single tree - DBScan



**Figure 4:** Results of the Branch and bound algorithm on the variant tables produced by the two human annotators and the DBScan clustering. The parsimony values are, for the first annotator, 104; for the second, 135; for the DBScan results, 292.

fr. 111.

If the results obtained from the automatic clustering of variants differ from the expertise of the human annotators, on the other hand the genealogical results do not necessarily contradict previous expertise, and seem to remain significant up to a point. Deciding whether the human annotators or the clustering have provided the most useful data remains to be fully assessed, but current results indicate that, even if the clustering differs in terms of classification of readings from human annotation (as shown by the mean Adjusted Rand Index), the results obtained through stemmatological analysis still remain indicative.

## 4. Conclusion

We've designed a tool that can facilitate the study of unilingual and multilingual variant traditions, and its stemmatological analysis. The results obtained through clustering and stemmatological analysis tend to show that an automatic clustering of variants and the application of a phylogenetic algorithm can quickly produce results that are indicative of the genealogy of witnesses, and comparable to some extent to a traditional human expertise.

Yet, in the current state, many aspects of the processing workflow could still be improved, especially if the results of the automatic alignment and collation were to be used for scholarly editing purposes.

As far as segmentation is concerned, the delimiter tokens are those that reliably identify coherent grammatical segments (relative pronouns, coordinating conjunctions between clauses, and subordinating conjunctions, etc.). This is a problem for all sentences that begin with personal pronouns, articles, or nouns and not with the markers listed above. Therefore, we need to produce more segmented data to be able to distinguish all these complex cases.

As for the alignment step, we show that LaBSE's sentence embeddings model works surprisingly well on medieval language states. Still, there is room for improvement and some fine-tuning of LaBSE model on aligned medieval texts should improve the results. Variant clustering should also benefit from this fine-tuning. However, it will be a complex and time-consuming task, because of the required data, that is aligned texts: this step is planned for the future. Another important step will be microscopic alignment, which involves further work on modeling the alignment process, because it differs quite significantly from the alignment of monolingual texts, particularly with regard to the variations in the word order in the target sentence and in the source sentence. The micro-alignment could allow also more precise results in the stemmatological step.

## Materials and code availability

The tool for alignment and collation is freely available on Zenodo: [10.5281/zenodo.12732533](https://zenodo.org/record/12732533) and is maintained on github: <https://github.com/ProMeText/Aquilign/>. The result data can be found in the `results_dir` directory. The segmenter train corpus is available in the subfolder: `data/tokenisation`. The data and scripts for clustering and stemmatological analysis are available on a specific Zenodo repository: [10.5281/zenodo.12728282](https://zenodo.org/record/12728282).



## Acknowledgments

Co-funded by the *Agence Nationale de la Recherche* under the Équipex Bibliissima+ (ANR, 21-ESRE-0005).



Co-funded by the European Union (ERC, LostMA, 101117408). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

The authors additionally wish to thank Florian Cafiero for his thoughts on this research.

## References

- [1] M. Artetxe and H. Schwenk. “Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond”. In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 597–610. DOI: 10.1162/tacl\\_a\\_00288.
- [2] L. Bambaci, F. Boschetti, and R. Del Gratta. “Qohelet Euporia: A Domain Specific Language to Annotate Multilingual Variant Readings”. In: *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*. 2018, pp. 266–269. DOI: 10.1109/cist.2018.596332.
- [3] D. J. Birnbaum and H. M. Eckhoff. “Machine-Assisted Multilingual Alignment of the Old Church Slavonic Codex Suprasliensis”. In: *V Zeleni Drželi Zeleni Breg*. Ed. by S. M. Dickey and M. R. Lauersdorf. Indiana: Bloomington, 2018, pp. 1–15. URL: <https://www.researchgate.net/profile/Nada-Sabec/publication/371120502%5C%5FToward%5C%5FLess%5C%5FFormal%5C%5FWays%5C%5Fof%5C%5FAddressing%5C%5Fthe%5C%5FOther%5C%5Fin%5C%5FSlovene/links/6473967b59d5ad5f9c803dac/Toward-Less-Formal-Ways-of-Addressing-the-Other-in-Slovene.pdf>.
- [4] L. Brun. *Arlima - Archives de littérature du Moyen Âge*. Ottawa, 2005. URL: <https://www.arlima.net/>.
- [5] L. Cadioli, ed. *Lancello: versione italiana inedita del Lancelot en prose*. Firenze: Edizioni del Galluzzo, 2016.
- [6] L. Cadioli. “Le Lancello: italien”. In: *La matière arthurienne tardive en Europe, 1270-1530 = Late Arthurian Tradition in Europe*. Rennes: Presses universitaires de Rennes, 2020, pp. 501–510.
- [7] F. Cafiero and J.-B. Camps. “Stemmatology: An R Package for the Computer-Assisted Analysis of Textual Traditions”. In: *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities CRH-2, 25-26 January 2018, Vienna, Austria*, ed. Andrew U. Frank; Christine Ivanovic; Francesco Mambrini; Marco Passarotti; Caroline Sporleder (2018).
- [8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. “Spanish Pre-Trained BERT Model and Evaluation Data”. In: *PML4DC at ICLR 2020*. 2020, pp. 1–9. URL: <https://arxiv.org/abs/2308.02976>.

- [9] A. Contreras Martín and H. L. Sharrer, eds. *Lanzarote del Lago*. Alcalá de Henares: Centro de estudios cervantinos, 2006.
- [10] M. T. Elbatta and W. M. Ashour. “A dynamic method for discovering density varied clusters”. In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 6.1.1 (2013), pp. 123–134.
- [11] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In: *kdd*. Vol. 96-34. 1996, pp. 226–231.
- [12] M. D. Hendy and D. Penny. “Branch and bound algorithms to determine minimal evolutionary trees”. In: *Mathematical biosciences* 59.2.2 (1982), pp. 277–290.
- [13] Irht. *Jonas: Répertoire des textes et des manuscrits médiévaux d’oc et d’oïl*. Paris et Orléans, 2016. URL: <http://jonas.irht.cnrs.fr/>.
- [14] O. Kraif. *Adaptative Bilingual Aligning Using Multilingual Sentence Embedding*. preprint. arXiv, 2024. URL: <http://arxiv.org/abs/2403.11921>.
- [15] L. Liu and M. Zhu. “Bertalign: Improved Word Embedding-Based Sentence Alignment for Chinese–English Parallel Corpora of Literary Texts”. In: *Digital Scholarship in the Humanities* 38.2.2 (2023), pp. 621–634. DOI: 10.1093/lc/fqac089.
- [16] C. Meinecke, D. J. Wrisley, and S. Jänicke. *Automated Alignment of Medieval Text Versions Based on Word Embeddings*. preprint. Open Science Framework, 2020. DOI: 10.31219/osf.io/tah3y.
- [17] A. Micha, ed. *Lancelot: roman en prose du XIIIe siècle. Tome II*. Genève: Droz, 1978.
- [18] A. Micha, ed. *Lancelot: roman en prose du XIIIe siècle. Tome IV*. Genève: Droz, 1979.
- [19] M. Sánchez Sánchez and C. Domínguez Cintas. “El banco de datos de la RAE: CREA y CORDE”. In: *Per Abbat: boletín filológico de actualización académica y didáctica* 2 (2007), pp. 137–148. URL: <https://dialnet.unirioja.es/servlet/articulo?codigo=2210249>.
- [20] G. Scala. “La Tradizione Manoscritta Del ”Livre Du Gouvernement Des Roys et Des Princes” Di Henri de Gauchy. Studio Filologico e Saggio Di Edizione”. PhD thesis. University of Zurich, 2021. DOI: 10.5167/uzh-202620.
- [21] S. Schweter. *Europeana BERT and ELECTRA models*. 2020. DOI: 10.5281/zenodo.4275044.
- [22] H. O. Sommer, ed. *The Vulgate Version of the Arthurian Romances*. Vol. 4, Le Livre de Lancelot del Lac, part II. Washington: Carnegie Institution of Washington, 1911.
- [23] H. O. Sommer, ed. *The Vulgate Version of the Arthurian Romances*. Vol. 5, Le Livre de Lancelot del Lac, part III. Washington: Carnegie Institution of Washington, 1912.
- [24] F. Steven R. “The Complete Medieval Dreambook. A Multilingual, Alphabetical Somnia Danielis Collation”. PhD thesis. University of Michigan, 1978.
- [25] J. Tiedemann. *Bitext Alignment*. 1st ed. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2011. URL: <http://gen.lib.rus.ec/book/index.php?md5=7fb8c6d1d4e8924f79a03026e23b6517>.

- [26] J. Tronch. “Displaying Textual and Translational Variants in a Hypertextual and Multilingual Edition of Shakespeare’s Multi-text Plays”. In: *Early Modern Studies after the Digital Turn*. Ed. by L. Estill, D. K. Jakacki, and M. Ullyot. Medieval and Renaissance Texts and Studies 502. Toronto, Ontario: Iter Press, 2016, pp. 92–116.
- [27] D. Witschard, I. Jusufi, R. M. Martins, K. Kucher, and A. Kerren. “Interactive Optimization of Embedding-Based Text Similarity Calculations”. In: *Information Visualization* 21.4 (2022). DOI: 10.1177/14738716221114372.
- [28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. v. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020, pp. 38–45. URL: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.

## Appendices

### A. Witnesses

- lancellotto: *Lancellotto*, italian witness, edited by Luca Cadioli (cf. bibliography). Firenze, Biblioteca della Fondazione Ezio Franceschini, 1 (last quarter of the 14<sup>th</sup> century).
- lanzarote: *Lanzarote*, castillan witness, edited by Antonio Contreras Martín and Harvey L. Sharrer (cf. bibliography). Madrid, Biblioteca Nacional de España, 9611 (16<sup>th</sup>, copy of a manuscript dated from 1414).
- sommer: edition of H. Oscar *Sommer* (cf. bibliography). Tomes IV and V: London, British Library, Additional 10293 (beginnings of the 14<sup>th</sup> century).
- micha: edition of Alexandre Micha (cf. bibliography). Tome II: Cambridge, Corpus Christi College Library, 45 (2<sup>nd</sup> half of the 13<sup>th</sup> century). Tome IV: Oxford, Bodleian Library, Rawlinson D. 899 (14<sup>th</sup> century).
- fr111: BnF, français 111 (15<sup>th</sup> century).
- fr333: BnF, français 333 (first quarter of the 14<sup>th</sup>).
- fr751: BnF, français 751 (2<sup>nd</sup> part of the 13<sup>th</sup> century).
- inc: BnF, RES-Y2-46 (vol. 1) and RES-Y2-47 (vol. 2) (Printed in Rouen [vol. 1, by Jean le Bourgeois] and Paris [vol. 2, by Jean Dupré] in 1488).

## B. Witnesses and segments of text

The *Lancellotto* is particularly fragmentary. Only the sections of the text corresponding to these fragments were studied.

According to Micha, it corresponds to the sections: II, XLVIII 29 – L 11; II, LXI 26 - LXIX 23; IV, LXX 1 - LXXXI 17. But, for the last, the *Lanzarote* omitted the text and starts at IV, LXXV, with a very reduced portion in LXXX 16 to 43. Thus, the studied section of text starts at IV, LXXV.

We indicated here the witnesses and the folios where the segments of the text can be found.

| Lancellotto   | Lanzarote  | Sommer      | Micha                   | proper name        | id    | status  |
|---------------|--|-------------|-------------------------|--------------------|-------|---|
| f. 222r-226v  | f. 203r-218v   | IV, 271-285 | II, XLVIII 29<br>– L 11 | suite Charrette I  | ii-48 | full  |
| f. 243r-254vb | f. 265r-286v to LXVIII, with reduced 8-12                  | IV, 327-362 | II, LXI 26-LXIX 23      | suite Charrette II | ii-61 | without the end in the <i>Lanzarote</i>                   |
| f. 254vb-297v | f. 286v-320v IV, LXXV-LXXX, with LXXX 16 à 43 very reduced | V, 3-143    | IV, LXX 1-LXXXI 17      | part of Agravain   | iv-75 | without the beginning and the end in the <i>Lanzarote</i> |

These passages correspond, in the other French witnesses, to:

| BnF fr. 751 | BnF fr. 333 | BnF fr. 111 | incunable   | id    |
|-------------|-------------|-------------|---|-------|
| 227va-232va |             | 138vb-141vb | Diii.v-Dviii.r  | ii-48 |
| 248va-259va |             | 151rb-157vb | Fiiii.r-Gv.v  | ii-61 |
| 270vb-291rb | 23rb-47va   | 170va-183va | Iviii.v- end vol. 1 (Liii.v) + vol. 2 beginning to Bv.v | iv-75 |

We can note that, as we indicated in the introduction, the fr. 333 only contains the third studied segment and that the text corresponding to the incunabula is contained by two different volumes.

## C. Example of missed section

As example of missed sections of text that can have an impact on the alignment process, we can take the fight in the episode of the hawk (corresponding to the segment ii-61-2).

In Micha's edition, the text is the following (the whole episode is given by all the other studied witnesses):

... tant qu'il vienent en une valee. Et lors li mostre un poi en sus del chemin a senestre la loge dont li chevaliers issi, dont ele se plaint. « Venés avant, fet il, seurement, et se vos veez l'esprevier, si le prenés, ja por nului nel laissiés. Et je vos creant loialment que je le vos garantirai a mon pooir contre tos cels qui le voldront contredire. Et se li espreviens n'i est, si me mostrez le chevalier qui le vos toli et jel vos ferai amender tot a vostre volenté. — Sire, fet ele, de Dieu aiés vos bone aventure ! Mais je voldroie miels que vos le me poissiés rendre a pes que a guerre. — Par Dieu, fet il, se ie ne le puis avoir par debonaireté, si l'avrai je par force. » Lors sont venu a la loge ; si entre mesure Yvain tos premiers et la damoisele après. Et mesure Yvain ne salue nul de cels de laiens, ains dist si haut que tuit le porent oir : « Damoisele, venés avant et si prenés vostre esprevier, se vos saiens le poez veoir ; si l'enportez ausi a droit com il en fu portez a tort. — Sire, fet ele, volentiers, ausi le voi je la. » Et ele vient a une perche ou il seoit, si li deslie les giés let l'en volt porter, quant uns chevaliers saut avant qui li dist : « Fuiés. damoisele ! Ne le remués, que par mon chief vos n'en portérés point. Et de tant com vos i estes retornee, avés vos del tot vos pas perdus, que vos ne l'enporterés n'en l'une main n'en l'autre ; et se vos volés avoir oisel, si querés autre, kar a cestui ne vos deduirés vos jamés. — Laissiés li, dans chevaliers, fet mesure Yvain, qu'ele l'enportera, et se vos li volés fere force, vos en serés tart al repentir. — Comment ? fet cil. Estes vos ci venus por le deffendre ? — Ce verrois vos bien, fet mesure Yvain, se vos ti tolés. » Et cil giete maintenant le main por le tolr ; et mesure Yvain a trait l'espee et li dist qu'il li coupera le bras, s'il toche plus ne a lui ne a la damoisele. « Voire, fet cil, par mon chief mal le deïstes ! » Lors cort a son hialme et le met en sa teste, et il estoit molt bien armés de tote autre armeure. Maintenant saut en son cheval, kar tos estoit prest, et prent son escu et son glaive et dist a mon seignor Yvain qu'il se gart de lui ; si li laisse corre, son glaive aloigné et mesure Yvain a lui ; si s'entredont si grans cops sor les escus qu'il les font fendre et percier et les haubers desmaillier et derompre ; si se metent es chars nues les glaives trenchans, si s'entrehurtent des escus et des cors et des visages, si s'entreportent a terre tot enferré. Mesure Yvain est navrés el costé destre et li chevaliers fu ferus par mi le cors si durement qu'il n'a pooir de soi relever de la ou il gist. Et mesure Yvain se redrece a tot le tronçon qui demi li est el costé ; si trait l'espee et s'apareille d'assaillir le chevalier qui le meillor cop li a doné qu'il receust pieça ; et il le cuide tot prest trover de deffendre, si voitqu'il ne se remue, et lors li cort sus et li esrache le hialme de la teste et dist qu'il li coupera le chief sans arest, s'il ne se tient por outré. Et cil parole a grant paine com cil qui molt estoit bleciés, si crie merci et dist : « Ha, frans chevaliers, ne m'ocie mie, mais laisse moi vivretant que j'aie mon Salveor receu, kar je sai bien que je sui navrés a mort ; si vos pri por Dieu que vos alés ci pres desus cest tertre querre un saint home mire qui i maint, et li faites avec lui apporter a corpus Domini. Et il dist que si fera il volentiers ; si commande la damoisele qu'ele s'en aut et ele le fet. Mais ele fet assés greignor duel que devant, kar ele voit .I. chevalier ocis et .I. autre navré, et por si petit d'acheison. Et mesure Yvain vet querre l'ermite, ensi com li chevaliers li ot dit, et li amaine. Et quant il fu revenus arrieres, si trueve iluec .I. escuier et une damoisele qui estoit amie al chevalier et faisoit le greignor duel del monde. Et quant li chevaliers fu confés et il ot receu son Salveor, si le coucha l'en en la loge. Et mesure Yvain s'en vet avec l'ermite et enmaine son cheval en destre, kar a cheval n'i alast il mie delés si haut saintuaire comme Nostre Salveor. Quant il furent venu a l'ermitage que l'en apeloit l'ermitage del Mont, si desarment troi frere qui laiens estoient mon seignor Yvain, et il en i avoit un qui molt savoit de plaies garir ; si s'entremist de mon seignor Yvain et s'en prist garde erraument et il osta le tronçon qu'il avoit el costé et s'estanchaé a sainier ; et de cele plaie demora mesure Yvain .XV. jors laiens.

In the *Lanzarote*, the episode is really reduced, since it presents only the following text:

... y andubieron tanto que llegaron ala Ramada ado estaua el cauallero y la donzella se lo mostro don yban quando le vio dixole señor cauallero yo vos Ruego que dedes su gauilan aesta donzella quele tomastes o sino enla vatalla sodes el gauilan no se lo dare yo dixo el Cauallero mas dela batalla presto so e luego se dexaron correr el vno contrael otro y el cauallero quebro su lança en don yban y don yban lo firio tan de Reçio quelo derribo en tierra todo atordido que no se pudo leuantar e don yban desçendio porle cortar la cabeza mas el le Rogo quele perdonase y que faria todo su mandado y don yban lo dexo y el Cauallero selo agradeşcio mucho y dio luego el gauilan ala donzella y ella se fue muy alegre conel y don yban se fue ensu demanda

This example shows the need to define what a collatable example is. An abstract is not really alignable with its source, and we need to go through another processing step and not try alignment. This shows the need to produce solid textual relationship typologies before making text processing decisions.

## D. Example of alignment table

Table 11: Example of good quality alignment taken from ii-48 fragment (selected witnesses, no correction performed), with the BERT-based segmentation. Segments are separated with pipe “|” characters.

| <b>micha</b>  | <b>fr751</b>   | <b>inc</b>  | <b>lanzarote</b>  | <b>lancellotto</b>                                      |
|---|--|---|---|---|
| mais il voit l’eve noire<br>et parfonde  et si peril-<br>luse | Mais il uoit leue<br>si parfonde et si<br>perilleuse | mes il uoit<br>la riuiere<br>parfonde et<br>dangereuse a<br>passer. | Mas el agua<br>hera muy fonda<br>e peligrosa e<br>negra e bien<br>cuidaua morir | ma e vede<br>l’acqua nera<br>e profunda e<br>perigliosa |
| qu’il cuide bien noier,                                       | que il cuide bien<br>perir                           | et scait bien   |   | che crede bene<br>morire                                |
| s’il se met dedens;   | se il se met<br>dedans.                              | sil entre de-<br>dens qu il se<br>met en peril de<br>mort.          | si se y Metiese   | s’elli si mette di<br>dentro;                           |
| et d’autre part il voit<br>cele                               | et dautre part il<br>uoit cele                       | Et daultre part il<br>uoit celle                                    | e dela otra parte<br>veya la donçella   | e d’altra parte e<br>vede colei                         |
| qui si durement crie<br>merci;                                | qui si docemet li<br>prie [mer]ci.                   | qui si piteuse-<br>ment fui crie<br>mercy.                          | que muy afin-<br>cada mente le<br>pedia merçed                                  | che si dura-<br>mente gli grida<br>mercé,               |
| si l’em prent tels pitiés                                     | Si len prent tes<br>pitiez                           | Si luy en prent<br>telle pitie                                      | e ovo tal piedad<br>della   | si ne gli prende<br>tale piatà                          |
| qu’il en laisse totes<br>poors                                | quil en laisse<br>totes paours.                      | quil en laisse<br>toute paour                                       | que le fiço todo<br>el miedo perder   | che ne lascia<br>tutte paure                            |

|  |   |   |  |  |
|--|---|---|--|--|
| et fet le signe de la<br>crois en mi son vis,  | Si fait le signe<br>de la crois enmi<br>son uis.  | et fait le signe<br>de la crois<br>deuant son<br>uisaige  | e fiço la señal<br>dela cruz sobre<br>si   | e si fa el segno<br>della santa<br>croce nel milu-<br>ogo del suo<br>viso,   |
| puis embrace l'escu  et<br>broche le cheval des<br>esperons  et se fiert en<br>l'ève.  Et li chevals<br>fu fors, si commence a<br>noer | puis embrace<br>lescu.  et<br>broche le<br>cheual des<br>esperons et se<br>fiert en leue tot<br>errantment.  Et<br>li cheuaus fu<br>fors si comance<br>a noer | puy embrache<br>son escu  et<br>fiert son cheual<br>des esperens<br> et se lanche<br>dedens leue<br>tout erraument:<br> et le cheual<br>commence a<br>noer                  | e enbraço el<br>escudo  e firio<br>al cauallo de<br>las espuelas<br> e lanço se<br>en el agua<br>e el cauallo<br>començo de<br>nadar luego | poscia imbrac-<br>cia lo scudo<br> e broca il<br>cavallo degli<br>sproni,  si fiede<br>nell'acqua is-<br>nellamente. El<br>cavallo fu forte,<br>incominciò a<br>notare |
| si tost com il ot terre<br>perdue,   | si tost com il ot<br>terre perdue   | si tost  quil eust<br>terre perdue.   | e perdio tierra<br>en tal manera   | si tosto com'elli<br>ebbe terra per-<br>duta,  |
| si l'enporte d'autre<br>part de la rive a<br>quelque paine,  mais<br>ançois ot beu de l'ève<br>li uns et li autre,                     | si lenporte<br>dautre part la<br>riue a quelque<br>painne.  | et sen passe le<br>cheual iusques<br>de laultre part<br>de leue  mes<br>ce fust a grant<br>paine  car ains<br>quilz feussent<br>passes beurent<br>de leue lun et<br>laultre | que paso dela<br>otra parte del<br>agua  mas ante<br>beuio el cauallo<br>del agua  | si ne l portò<br>d'altra parte<br>della riva a<br>qualche pena,<br> ma inanzi<br>furono amen-<br>duni tutti<br>molli,  |
| et se li chevals ne fust<br>si buens,  noié fuissent<br>ambedui,   |   | et se le cheual<br>neust este<br>fort et bon ilz<br>feussent noyes<br>et lun et laultre   | e si el caballo<br>tan bueno no<br>fuera sin dubda<br> el moriera enel<br>agua   | e se l cavallo<br>non fosse<br>stato si forte,<br>anegati fossono<br>amenduni,   |
| kar li chevaliers estoit<br>pesans por les armes   |   | car le cheualier<br>estoit fort pe-<br>sant pour les<br>armes   | e boores  porel<br>peso delas ar-<br>mas   | ché l cavaliere<br>era pesante per<br>l'arme   |
| qu'il avoit vestues.   |   | quil auoit<br>uestues.  | que llebaua  | ch'elli avea<br>vestite.   |

|  |                    |  |  |   |
|--|--------------------|--|--|---|
| Quant il fu de l'autre part de l'eve,  si ne descendi pas, | Si ne descendi pas | Quant il fut de lautre part de leaue  si ne descendist oncques | e desque fue dela otra Parte del agua dexo se correr aquellos que la donçella tenian | Quand'e fu da l'altra parte dell'acqua,  si non discende passo, |
|--|--------------------|--|--|---|



## E. Readings and groups of witnesses

### E.1. Specific readings

In some sections, every witness has its own reading (fragment ii-61-1):

| micha   | sommer   | fr751  | fr111   | inc                                  | lanzarote                                       | lancellotto   |
|---|--|--|---|--------------------------------------|---|---|
| Puis traist hors de sa char la piece de l'espee qui dedens estoit | puis traist hors de sa quisse le piece de lespee | Puis trait la piece de lespee qui en sa char estoit. | puis trait hors de sa char la piece de lespee | apres trait hors de sa cuisse lespee | e despues tiro la pieça dela espada desu pierna | Poscia trae fuori di sua carne la spada e l'apicca: |

### E.2. Examples of readings specific to two main groups

The two main groups can be established from common omissions, for example (fragment ii-48):

| micha   | sommer | fr751   | fr111  | inc | lanzarote | lancellotto   |
|---|--------|---|--|-----|-----------|---|
| si le font desarmer, ou il volsist ou non, et il le fist a envis, |        | Si le font desarmer uossist ou non. et il le fait mout enuiz. | si le font desarmer uoulsist il ou non. et il le fist a peine. |     |           | si l fanno disarmare o volesse O non, ed elli il fece ad invidia, |

We can also identify common variants that systematically oppose group 1 to group 2. In the following table (fragment ii-48), the two variants oppose *shield* to *weapons*, *castle* to *forest*.

| micha                                  | sommer                                     | fr751                                  | fr111                                  | inc                                       | lanzarote   | lancellotto                                |
|--|--|--|--|---|---|--|
| mes il ne portoit mie tel <b>escu</b>  | mais il ne portoit mie tels <b>armes</b>   | mais il ne portoit mie tel <b>escu</b> | mais il ne portoit pas tel <b>escu</b> | mes il ne portoit mye telles <b>armes</b> | mas no traia el tales <b>armas</b>                  | ma e' non porta mica tale <b>scudo</b>     |
| Al <b>chastel</b> , fet ele, de Floego | En la <b>forest</b> fait elle de floregas. | Au <b>chastel</b> fait ele de floego.  | ou <b>chastel</b> de fleago fait ella  | En la <b>forest</b> de florega:           | enla <b>forest</b> de donseglorega dixo la donçella | Al <b>castello</b> , diss'ella, di Fleego, |

### E.3. Oscillation of *Lanzarote* between two groups

*Lanzarote* belongs to the second group, but presents sometimes readings from the first one, or even a mixed reading, as we can see in the following table (fragment ii-48):

| lanzarote  | micha  | sommer   |
|--|--|--|
| e el fue alla por aluergar e fallo <b>ala puerta dos</b> frailes | et il torne cele part por herbergier. Et quant il i vint, si trueve <b>a la porte IIII.</b> des freres | si torne cele part pour herbergier. Et quant il vint la si trouua <b>.ii.</b> freres |

The first group, with Micha, is characterized by the mention of the *porte* and the presence of four *freres* whereas the second one, the Sommer's one, is characterized by the absence of the mention of the *porte* and the presence of only two *freres*. The text in the *Lanzarote* presents the mention of a *porte* and only two *freres*.

### E.4. Example of innovation in *Lanzarote*

*Lanzarote* presents some specific innovations compared to the other witnesses in the following table (fragment ii-48):

| micha  | lanzarote  |
|--|--|
| puis voit delés les chevaliers une tombe, la plus riche  | e el asi Catando vio dentro enlos arcos vna muy Rica tumba   |
| qui onques fust fete par home, kar ele ert tote d'or fin a chieres pierres precioses qui molt valoient miels d'un grant roialme. Se la tumba fu de grant bialté, nient ne monte la bialté envers la richece dont ele estoit et avec ce estoit ele la greignor que Lancelos eust onques veue: | qual nunca tal viera el ni ome del mundo que hera toda de oro fino e de piedras preçiosas que estaua toda labrada de diuersas maneras e de muy muchas cosas de figuras e de otras cosas que tal tumba de tal manera nunca viera ni ome del mundo que la non ouiese visto no podria asmar la fermosura que enella auia de tales cosas como enella estauan |
| si se merueille molt qui puet estre li princes   | e don lançarote se marauillo e dixo entresi mesmo Quien podria ser el príncipe   |

### E.5. Opposition between *Lancellotto* and Micha against BnF fr. 111

We can find a multitude of readings that oppose the group *Lancellotto*/Micha against BnF fr. 111 (fragment ii-61-1), such as the ones in the following table:

| lancelotto   | micha   | fr111   |
|--|---|---|
| se noi abiamo per miscre-<br>denza fallato <b>qua in adi-<br/>etro</b>               | se nos avons par mescre-<br>ance foloié <b>ça en arieres</b>                | se nous auons foloye par<br>mescreance  |
| ... che trastutti cristinia-<br>vano, si confessò, <b>udendo</b><br>tutto el popolo, | ... que tuit se crestienoient,<br>si reconuit <b>oiant</b> tot le<br>pueple | ... que tous se<br>c[re]stienneroit. si cogneut<br><b>deuant</b> tout le peuple |

Some variants are more important (fragment ii-48):

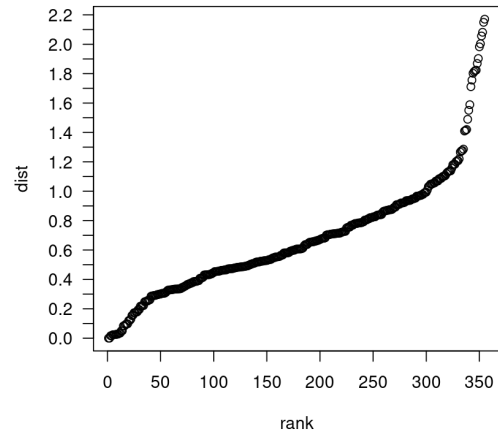
| lancelotto  | micha  | fr111   |
|---|--|---|
| <b>e quelli iscioglie im-<br/>mantanente sua caccia-<br/>gione e la</b> dona a suo<br>compagnone per portar-<br>nela, | <b>et cil se destrosse main-<br/>tenant de sa venoison,<br/>si la</b> baille a son com-<br>paignon por porter, | et cil baille sa uenoison a<br>porter a son compaignon. |

## E.6. Example of innovation in *Lancelotto*

*Lancelotto* presents in some passages innovations (fragment ii-61-2):

| micha   | lancelotto  |
|---|---|
| et de ceu est il tos esbahis. Après re-<br>garde la pucele, si se merueille plus as-<br>sés de sa bialté que del vaissel, | e di ciò si maraviglia egli molto, si nè<br>molto isbigotito.   |
|   | Ma di cosa che veggia non si mar-<br>aviglia egli niente inverso della<br>trasgran biltà della damigella che<br>tanto gli sembra esser bella ch'a pena<br>l'osa riguardare, ché più gli è aviso<br>che sua faccia risprenda che sole, e<br>suoi occhi rilucenti e sua capellatura<br>che gli sembra ad essere di fine oro; ve<br>di persona si, come dice in suo cuore, |
| kar onques mes ne vit il feme   | non ne vide mai niuna   |

## F. DMDbscan



**Figure 5:** Minimum distances to the MinPts nearest points, sorted by ascending order. Two different densities levels are possible, the first one breaking around 0.2 distance, and the other around 1.2.