Make Love or War? Monitoring the Thematic Evolution of Medieval French Narratives

Jean-Baptiste Camps, Nicolas Baumard, Pierre-Carl Langlais, Olivier Morin, Thibault Clérice, Jade Norindr

Make Love or War? Monitoring the Thematic Evolution of Medieval French Narratives

Jean-Baptiste Camps[1,*], Nicolas Baumard[2], Pierre-Carl Langlais[3], Olivier Morin[2], Thibault Clérice[4] and Jade Norindr[1]

[1]*École nationale des chartes - Université PSL, 65 rue de Richelieu, Paris, 75012, France*
[2]*École normale supérieure - Université PSL, 29 rue d'Ulm, Paris, 75005, France*
[3]*OpSci, 3 rue de Milan, Paris, 75009, France*
[4]*ALMAnaCH - Inria, 2 Rue Simone IFF, 75012 Paris*

Abstract

In this paper, we test a famous conjecture in literary history put forward by Seignobos and de Rougemont according to which the French central medieval period (12-13th centuries) is characterized by an important increase in the cultural importance of love. To do that, we focus on the large and culturally important body of manuscripts containing medieval French long narrative fictions, in particular epics (*chansons de geste*, of the Matter of France) and romances (chiefly *romans* on the Matters of Britain and of Rome), both in verse and in prose, from the 12th to the 15th century. We introduce the largest available corpus of these texts, the *Corpus of Medieval French Epics and Romances*, composed of digitised manuscripts drawn from *Gallica*, and processed through layout analysis and handwritten text recognition. We then use semantic representations based on embeddings to monitor the place given to love and violence in this corpus, through time. We observe that themes (such as the relation between love and death) and emblematic works well identified by literary history do indeed play a central part in the representation of love in the corpus, but our modelling also points to the characteristic nature of more overlooked works. Variation in time seems to show that there is indeed an phase of expansion of love in these fictions, in the 13th and early 14th century, followed by a period of contraction, that seem to correlate with the Crisis of the Late Middle Ages.

Keywords

Medieval French Literature, Cultural Evolution, History of Emotions, Document Analysis and Recognition, HTR, Word and Document Embedding

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France
*Corresponding author.
Jean-Baptiste.Camps@chartes.psl.eu (J. Camps); nbaumard@gmail.com (N. Baumard); pierrecarl.langlais@gmail.com (P. Langlais); olivier.morin@ens.psl.eu (O. Morin); thibault.clerice@inria.fr (T. Clérice); jade.norindr@chartes.psl.eu (J. Norindr)
https://www.chartes.psl.eu/fr/jean-baptiste-camps (J. Camps); https://nicolasbaumards.org/ (N. Baumard); https://sites.google.com/site/sitedoliviermorin/ (O. Morin); https://github.com/ponteineptique (T. Clérice)
0000-0003-0385-7037 (J. Camps); 0000-0002-1439-9150 (N. Baumard); 0000-0002-6216-1307 (O. Morin); 0000-0003-1852-9204 (T. Clérice)

1. Introduction: love and war in medieval French narratives

1.1. Love, a 'medieval invention'...

Love – or more precisely love in literature – is sometimes depicted as "a medieval invention", or rather, to quote the exact formulation of this phrase that goes back to the historian Charles Seignobos, "Love dates from the 12th century" [1][1]. What Seignobos meant is not that there was in Antiquity no conception of love, but he differentiates the antique notion of *Eros*, interpreted as sexual *desire*, at least for males (he admits the idea of love-as-respectful-devotion in Antique women) from the modern (and in his mind, Western only) notion of reciprocal love, that he defines as "a new feeling of respect and reciprocal admiration, supposing equality between the two sexes" [1]. This conception would find its origins in the 12th century "courtly love". The cultural movement of courtly love, the *fin'amor*, started in Southern France in the lyrical poetry of the *troubadours* around the start of the 12th century as far as we can judge, and knew a spectacular expansion spanning Western Europe and fecundating other types of literary productions, such as the new form of the *romans* (romance), and even epic narrative forms, until then more preoccupied with violence, lineage and feudal values, such as the *chansons de geste*, eventually blurring the frontier between the two genres.

In his masterwork, *L'Amour et l'Occident* (*Love and the Western World*) [2], Denis de Rougemont makes of the myth of *Tristan et Iseut* the archetype and the most emblematic early representative of the love-as-passion in Western culture, of which some significant features are the link between love and death, the unsatisfied, frustrated or fatal issue of a yet reciprocal sentiment, the transgression of moral norms or social duties and, *in fine*, the adulterous nature of the relationship.

In addition, de Rougemont also draws a correspondence between two apparently antagonistic themes: love and war. Noting the use of military vocabulary in the depiction of the conquest of the loved lady, to which the lover must lay siege, after he has been struck by the arrows of Love, he argues that both (courtly) love and (gallant) war are realised in the same chivalric ideals ("*La chevalerie, loi de l'amour et de la guerre*": chivalry, law of love and war).

De Rougemont's work has received some criticism, because there is no unique definition of love in the Middle Ages, and that the refined love (*fin'amor*) of the lyrical poet has substantial differences with the passionate 'crazy' love of Tristan and Iseut (*fol'amor*) [3], and no medieval story fits all aspects of courtly love as a deliberate choice (henceforth, adulterous in nature) and a reciprocal sentiment: the troubadours most often complain of the disdain or excessive pride of the lady they love (hence, not always recriprocal), while Tristan and Iseut magical-filter-induced love does not perfectly fit the idea of a deliberate choice. Despite these reservation, and as far as the surviving documentation allows us to see, the 12th and 13th centuries saw an explosion of fictional love stories, both of known writers such as Beroul, Chrétien de Troyes or Marie de France for instance, or in the many anonymous works of this period, such as *Floire and Blancheflore* or *Aucassin and Nicolette* [4]. If these were written first in Occitan and French – be it Continental French, Anglo-French, or Franco-Italian –, equivalent in other Western European

[1]The more catchy "Love is a modern invention", Seignobos explains, was a corrupted version of what he said to a lady, who told the journalist Gustave Téry, who told a colleague, Henri Bellamy, who in turn published it in the *Quotidien*, spurring Seignobos reply.

languages were soon to appear, perhaps with the exception of Spain where the first examples of symmetric and passionate narrative love story arrive later in the 14th century [4]. For instance, in addition to the French version of the *Tristan and Iseut* story by Beroul, Thomas of Britain, Marie de France and Chrétien de Troyes, we also see German (Gottfried von Strassburg), Italian (*Tristano Riccardiano*, *Tristano Veneto* and *Tristano Corsiniano*), English (*Sir Tristrem*) and Czech (*Tristam and Izalda*) versions, and slightly later in Spanish from the late 14th c. to the 16th [4]. As Morris notes: "As far as our surviving evidence takes us, there was an enormous explosion of interest in the subject shortly before 1100. An almost complete silence was followed by the beginning of love literature which challenged in quality and surpassed in volume that of any earlier civilization" [5, 4].

The 14th and 15th century constitutes a period somewhat less explored with respect to Medieval French literature, despite – or probably because of – a large body of prose works, very often of consequent dimensions, and sometimes with abundant surviving manuscript traditions. Many of these works were new versions of previous texts, such as new versions of *Tristan and Iseut* or *Floris and Blancheflore*, and they are accompanied by many seemingly new creations such as *Ponthus et Sidoine* or *Cleriadus et Meliadice*, while, in other European languages, the works of Chaucer or the Middle Dutch plays of *Esmoreit* or *Gloriant* feature an important dimension of reciprocal love as passion [4]. Yet, to some extent, it remains to be seen if, in Medieval narratives, the importance of the theme of love and passion actually increases or decreases in the literature of the Late Middle Ages.

1.2. ... in broader perspective

If the importance of the development of love in medieval Western culture from the 12th century onwards cannot be denied, research now tends to put it back in a context where similar increases happened elsewhere in Eurasia, for instance in the Arab world, India, Persia, China and Japan, as well as in the West in other periods of time, such as the the Greece of the first to third centuries AD (that saw the production of 'novels' such as *Leucippe and Clitophon*) [6].

The medievalist Georges Duby was perhaps the first to hypothesize that economic development might be the main driver in the rise of love in Western Europe [7]. Recently, Baumard et al. [6] argued that a 'higher level of economic development' (approached through measures such as GDP per capita, urbanisation rate, size of the largest city, ...) 'is strongly associated with a greater incidence of love in narrative fiction', in the Eurasian space, both in the Antiquity and during the Middle Ages and Early Modern period [6].

In line with these work, we thus test whether there are indeed shared trends between love and economic development in literary fictions, on the specific corpus that played a central cultural role in medieval Europe and spurred de Rougemont's analysis: medieval French long narrative fictions. We compare it to measures of economic development, based on data available data for GDP in medieval France [8, 9]. It is to be noted that, if both appear correlated, it won't necessarily mean that increase in GDP causes increase in taste for love in fiction, as both could be influenced by external factors that are not yet included in our analysis, such as, for instance, political stability, or the absence of major shocks (wars, pandemics,...).

To go beyond the received (and relatively cramped) literary canon, in terms of works, authors, genres and periods, and to proceed on the material basis of the reception and popularity

of the works, we build a large corpus of manuscripts of *chansons de geste* as well as verse and prose *romans*, from the 12th to the 15th century. By working on the basis of the surviving manuscripts, we hope to circumvent some biases due to, both, the romantic and scholarly reception of medieval works in the 19th-21th centuries, such as the overvaluation of works fitting contemporary aesthetic criteria (rather than popular during the Middle Ages), but we are subject to biases in the differential preservation and destruction of these works [10, 11].

By including both epics (*chansons de geste*) and *romans*, we will also be able to go beyond traditional genre definitions to monitor the importance of the theme of love, and its ability to cross generic boundaries.

Finally, to give context to the evolution of the theme of love, and to test the hypothesis of de Rougemont of a link between them, we will also follow the chivalric theme of violence and war.

2. Materials and Methods

2.1. Global design and justification

Our global experimental design is as follow:

1. gather a corpus as large as possible of manuscripts of medieval French epics and romances, through the harvesting of digitised manuscripts and their subsequent processing through a dedicated workflow using computer vision and natural language processing;
2. build a semantic representation of words and documents, based on a joint embedding, using doc2vec, and estimate its quality using literary knowledge;
3. compute scores for documents, based on cosine similarity in the joint embedding between them and the vectors of words for love, and for violence;
4. monitor their variation during the period;
5. compare the variation with historical knowledge on economic development: do they converge, diverge or seem unrelated?
6. (appendix) use top2vec, based on the doc2vec embedding, to look at the topics with high love or violence scores, to check that they are indeed related to courtly love, and to violence, by interpreting them in light of literary history.

As there is no unequivocal definition of what constitutes the theme of love or of violence, we choose to constitute them by concatenating the vectorised word representations of:

- the lexemes *aimer* and *amour*, and their inflectional and spelling variants that we could identify[2];
- the lexeme *ferir* (*frapper*, hit) and its inflectional and spelling variants that we could identify. *ferir* was chosen because it constitutes probably the most ubiquitous verb in descriptions of fights [12, p. 149].

[2]We do not target specifically physical love and sexuality (and important texts in that regard, such as the *Fabliaux*, are indeed not included in our corpus), though these themes might still have appeared, as well as forms of non romantic love (e.g., divine, familial, etc.). Yet, the results below show that our vector captures almost exclusively courtly love themes.

Table 1
Corpus size, for the full corpus (CMFER-*full*), and the selection used for the analysis, with limitation to the years 1200-1500 and to witnesses with a ratio of lines with estimated good or average HTR quality ≥ 0.78 (CMFER-*select*; see Appendix A)

Corpus	MSS	Witnesses	Word tokens
CMFER-*full*	265	409	38.5M
CMFER-*select*	203	370	36.4M

2.2. Corpus

The *Corpus of Medieval French Epics and Romances* introduced in this paper is, to our knowledge, the largest corpus of Medieval French created until now. Though still in early version and with only partial coverage of its final scope, it is as of now comprised of 265 manuscripts, and 410 text witnesses, for a total of 38.5 million word tokens. The deep learning based workflow for text acquisition from the digitised manuscripts images, as well as the subsequent ground-truth free quality evaluation of the results are depicted in Appendix A.

2.2.1. Scope

The goal of the corpus is to encompass every manuscript of medieval French long narrative works, that fall broadly in the category of *chansons de geste* (epics) and chivalric *romans* (romances), chiefly but not exclusively from the Matters of Rome or Britain, in verses, along with their *mises en prose* and native prose versions, as long as they are available in digitised form.

For this paper, the scope was limited to digitisations available as part of the *Gallica* digital library of the Bibliothèque nationale de France. In the near future, it will be expanded to include other sources (such as the *Bibliothèque Virtuelle des Manuscrits Médiévaux*)[3].

The inventory of works, texts and manuscripts (still ongoing) was made by collating a list of epics made by one author, data from the *OpenStemmata* repository [13], with the list published by Kestemont et al. [10]. Verifications were made by going back to the digital catalog of the BnF [14], and online databases *Jonas* and *Arlima* [15, 16]. In particular, data was enriched with links to available digitisations.

2.2.2. Corpus facts and figures

Main statistics about the corpus are presented in Table 1. Due to the unequal availability of digitised manuscripts (and of the underlying sources), as well as selection on the basis of handwritten text recognition (HTR) quality, the corpus is not chronologically balanced, and no regularisation was performed on this aspect (which in part supposedly reflects also variations in manuscript production and preservation).

[3]The version used for this paper is limited to a subset of 258 manuscripts, due to time and computing resources constraints. It will be expanded in the course of the following months, to encompass all digitised manuscripts from our list of 800 manuscripts containing relevant texts and kept in the *Bibliothèque nationale de France*. In the context of the preparation of this paper, we have focused on increasing the diversity of texts rather than including, say, all the very numerous manuscripts of the prose arthurian *Vulgate* cycle or the *Guiron le courtois*.

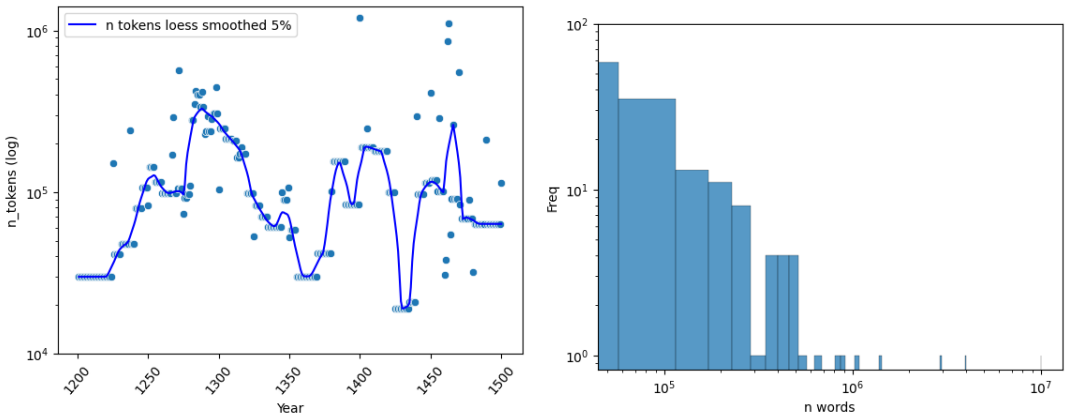


Figure 1: Variation in time of the number of tokens available in the corpora (tokens from witnesses with vague date ranges, e.g. 1201-1225, have been distributed equally all over the range)

The chronological distribution of the CMFER-*select* corpus (Figure 1, left) shows that substantial data is available for all the period envisioned. It also shows an almost continuous increase during the 13th century, followed by a very important decrease, hitting a lowest point in the years following the Black Death pandemic (associated with a significant population drop). It then increases again in the 15th century, with another notable drop during one of the worst periods of the Hundred Years' War, roughly the Armagnac–Burgundian Civil War (1407-1435). Even though biases in the availability of sources and choices made for the corpus are likely to be present, the correspondence with important historic events might be an argument for a form of representativeness of the corpus with respect to the medieval production, or perhaps with the inheritance of the Royal Library, established by Charles V the Wise in 1367 (the ancestor of the BnF, from which digitised manuscripts were obtained).

It is to be noted that the unequal distribution of tokens in time is not necessarily in itself a problem to estimate the average importance of love, as long as enough material is present throughout the period.

The distribution of number of tokens by work (Figure 1, right) shows, as expected, a very unequal distribution: works are most often found in a single witness, but a handful of texts have an abundant tradition, reflecting their enduring success. Some of the latter are very long or cyclical works, and in definitive, can amount for up to two orders of magnitude more tokens than most works: this is in particular the case of the prose *Tristan* (23 witnesses and 10M tokens, more than a fourth of the corpus), *Guiron le courtois* (9 witnesses and 4M tokens), *L'Estoire del Saint Graal* (20 witnesses and 3M tokens), or *Garin le Loherain* (11 witnesses and 1M tokens). We took the decision not to restrain the number of witnesses for a given text, because the aim of our analysis is to reflect the reception of the texts. In a context where books are expensive objects, commissioning the copy of a voluminous work is a significant choice.

Table 2
doc2vec hyperparameters (Train size in number of word tokens)

Method	Training size	Vector size	Window size	Min count	Sub-Sampling	Negative Sample	Epoch
dbow	36.4M	300	15	25	10^{-5}	5	5

2.3. Semantic representation of the words and documents

2.3.1. Model training

Given the level of lexical, spelling and abbreviative variation in the corpus, as well as the noise induced by the HTR process, and the current absence of subsequent normalisation such as lemmatisation, we are faced with an important amount of variant forms. To deal with this, we choose a method that is supposed to increase robustness to this type of variation, by creating a shared embedding of words, using word2vec [17], and documents, with doc2vec [18]. In addition, this allows us to use top2vec to extract topic vectors, to investigate the contexts in which our queried word vectors are used, to ascertain that they do, in fact, represent occurrences of courtly love or violence (see Appendix D).

Given the nature of our corpus, we are chiefly interested in several of the main claimed features of these embeddings, in particular the fact that they supposedly do not need stemming or lemmatisation, nor lists of stop words. In addition, some benchmarks have also found doc2vec to be the most efficient model over encoders, such as the Universal Sentence Encoder or BERT Sentence Transformer, [19] when used in contexts such as topic discovery. Last but not least, since our main goal is to interrogate the documents based on the importance of the semantic content related to the forms of the lexemes *amer*/*amour* and *ferir*, the advantage of using a combination of word2vec, doc2vec and top2vec is that it allows us to manipulate and interrogate shared representations of word, document and topic vectors.

Given the large size of the texts, they were sampled in 15 lines fragments (resulting in 334 060 fragments). The doc2vec model was trained with mostly default hyperparameters, with additionnal adjustements based on existing benchmarks, and the specifics of our corpus (Table 2). In particular, regarding the number of training epochs, previous studies on Doc2Vec found the optimal number of epochs for a fairly large corpus (in terms of document length and number of documents) to be relatively low: for a 4.5 million words corpus, Lau and Baldwin [20] found the optimal number of epochs to be 20, as opposed to 400 for a 0.5 million words corpus, and the minimum frequency of a word in the corpus for inclusion to be 5 instead of 1. Curiskis et al. [21] showed that for a dataset of approximately 7 000 documents of a mean length of 140 words, the optimal number of training epochs was 50. Since our corpus is closer to 40 million words and 300 000 samples, respectively one and two orders of magnitude larger, we retain the option of training for five epochs, with a minimum count of 25. In addition, we chose to use negative sampling instead of a hierarchical softmax step at the output layer because it proved both more efficient and yielding better quality vectors in existing benchmarks [22, 20], and chose a vocabulary based only on word 1-grams.

Training was made on a dedicated server, using 8 parallel workers.

2.3.2. Interrogation of the resulting vectors

Topics, texts and passages were then interrogated using the following methodology:

1. word vectors were interrogated on the basis of the lemmas 'amer' (*to love*) and 'amor' (subst. *love*), on one hand, and 'ferir' (*to hit*) on the other, to retrieve most similar words. Other forms (flexional, spelling, segmentation variants, or variant forms due to HTR noise) of the lemmas were then identified, and added to the request (e.g., 'amour', 'lamor', 'lamour', 'amors', 'amours', 'samor', 'damors', 'amo', 'amoit', 'lamoit', 'amee', etc.), iteratively, until the most similar stopped yielding forms of the lemmas (Appendix C).
2. those sets of words and their corresponding vectors were used to examine their direct environment, in terms of word vectors closest to them, as well as in terms of document vectors closest to them (both in cosine similarity), in order to establish the semantic contents and the nature of the works that they would retrieve, and to verify they were concerned with courtly love, and chivalric violence. In addition to this verification of the quality of the embedding, topic modelling was used more secondarily to look directly at the closest associated themes (Appendix D);
3. finally, those sets of word vectors were used to compute a love and a violence score (based on cosine similarity) for each document, and monitor the variation of this score through time. For this, the score for love and violence of all passages was retrieved, in order to calculate a yearly mean. This necessitated to distribute the passages chronologically based on their date or approximate dating: for instance, a passage in a manuscript dated to 1245 was assigned to the year 1245 with a weight of one; a passage dated to the last quarter of the 13th century was assigned to the years 1276 to 1300 with a weight of $\frac{1}{1300-1275} = 0.04$ for each year. The mean scores were then computed and plotted as a time serie, using local regression with the LOESS method (locally estimated scatterplot smoothing), with a smoothing coefficient of 0.15.

3. Results

3.1. Semantic environment of love and violence

In order to inspect the validity of the vectors of love and violence, and to establish to what specific kind of love or violence they referred, we looked at the contexts, through the interrogation of most similar word vectors, based on cosine similarity between our love and violence vectors (mean of word vectors that compose them) the vectors for each word in the model (Table 3). We completed it through topic modelling (Appendix D).

The words closest to the love vector exhibits a catalogue of courtly love vocabulary: in the designation of the lovers, in the expectations, languishing troubles and (metaphoric or not) death from love (as well as potential love quarrels); the use of feudal vocabulary (loyalty, feudal possession), the expression of feelings and its traditional metaphoric elements (fire, heart, ...), as well as love promises, desire and kisses. We also find courtly qualities (beauty, goodness, high social extraction), and their traditional incarnated opposites, be they jealous or simply not possessed of courtly qualities (the villein and their supposed vileness or boorishness).

Table 3
Word vectors most similar to the mean love and violence vectors (cosine distance); only the 40 most similar are shown here. The asterisk signifies forms with HTR mistakes

word	transl.	cosine dist.	word	transl.	similarity
amanz	*lover*	0.73	lifent	*splits his*	0.85
aamie	*friend/lover*	0.72	soncop	*his blow*	0.84
pramesse	*promise*	0.71	*lehaub	*the Hauberk*	0.82
mesmai	*I trouble myself*	0.71	enabat	*he slaughters*	0.82
dedame	*of lady*	0.7	litrence	*chops off his*	0.82
secroi	*if I believe*	0.7	desmailla	*broke the (chain)mail*	0.81
mocit	*kills me*	0.69	ronpi	*broke*	0.81
bonaire	*of good lineage; sweet*	0.69	elespee	*and the sword*	0.8
saisine	*possession*	0.69	fausa	*distort*	0.8
anemie	*enemy*	0.69	esdens	*in the teeth*	0.8
motroi	*grant me*	0.69	resailli	*jumped back*	0.79
lariens	*the thing*	0.68	fausart	*faussart*	0.79
chastoie	*castigate*	0.68	meschoisi	*fail to see*	0.79
auilain	*of peasant; vile*	0.68	abrandie	*he brandished*	0.79
beance	*desire*	0.68	*9sint	*chases*	0.79
auilenie	*of boorishness*	0.68	*lunet	*puts in his*	0.79
ialous	*jealous*	0.68	porfendi	*he cleaved*	0.78
mesprent	*inflames me*	0.67	*toide	*stiff*	0.78
*loiai	*loyal*	0.67	porfent	*he cleaves*	0.78
desplace	*moves* (phys.)	0.67	fandi	*he cleaved*	0.78
mescroit	*believes not*	0.67	tronco	*a slice*	0.78
*sabiante	*her beauty*	0.67	liualut	*earned him*	0.78
*loiauce	*loyalty*	0.66	copli	*him a blow*	0.77
maioie	*my joy*	0.66	ecercle	*on the hoop (of helmet)*	0.77
nuliour	*no day*	0.66	*lempai	*rams him*	0.77
cuerc	*heart*	0.66	delasele	*of the saddle*	0.77
*blasie	*blame*	0.66	lespiet	*the spear*	0.77
esmuet	*moves* (emot.)	0.66	*delbu	*of the trunk*	0.77
daigniez	*deign*	0.66	*porfe	*he cleaves*	0.77
languist	*languishes*	0.65	laubc	*the hauberk*	0.77
porpens	*thought, imagination*	0.65	*aleust	*he would have had him*	0.77
descort	*discord*	0.65	lasena	*slammed him*	0.77
abonte	*of goodness*	0.65	laubt	*the hauberk*	0.77
baisiers	*kisses*	0.65	enlescu	*on the shield*	0.77
decoit	*deceives*	0.65	esqu	*shield*	0.76
socirra	*will kill her/himself*	0.65	*renche	*he slices*	0.76
mesperance	*my hope*	0.65	liuaut	*earns him*	0.76
metient	*holds me*	0.65	lenpaint	*rams him*	0.76
enfeist	*made of it*	0.65	desrout	*he shatters*	0.76
mesfis	*I wrongly did*	0.64	elpis	*in the chest*	0.76

The words closest to the *ferir* (hit) vector form an even more compact vocabulary: it is about hitting one's opponent with offensive weapons in the teeth, chest or shield, breaking pieces of armour, slashing, cleaving, slicing, piercing through, throwing him off his horse, and ultimately killing him.

Given these results, we are satisfied that the word embedding offers a relevant representation of (courtly) love and violence (especially, chivalric combats). We then move on to examine the

document embedding, on the basis of these word vectors.

3.2. Document scores for love and violence

Document-level scores for love and violence were computed for each textual witness, by taking the mean score of all passages extracted from them. If we rank them accordingly (Table 4), we notice in both lists the importance of manuscripts dating to the 13th and the early 14th century. The list of witnesses closest to the love vector show the importance of courtly love stories, in a mix of works whose literary importance is often known and sometimes less so: the very famous *Lai de l'ombre*, for instance, is an archetypal courtly tale by Jean Renart, in which a knight seduces a woman that was refusing him, by gifting a ring to her reflection in a fountain. The list also contains several adventure and love romances centred on a couple, such as *Amadas et Ydoine*, *Floire et Blancheflor* (here in its 'aristocratic' version), *Cristal et Clarie*. Several of these works share an Ovidian inspiration, and narrative patterns typical of courtly love (such as the gift or exchange of rings). The works of Adenet le Roi also feature in good place, be it the courtly adventure romance of *Cléomadès*, or the *Berte aus grans piés*, in which he mixes epic sources with a fine description of the feelings and troubles of its chief female character. Some lesser known texts fit quite well in this list: the highest scoring one is the *Roman de la poire*, a text in-between romance and lyrical poetry, in which "the themes of courtesy are present, with sophistication and refinement pushed to the extreme" [23, our translation], that centers around the initially non reciprocated love between the narrator and a lady, communicating through lyrical poems, and benefiting from the mediation of allegories of Love, courtly virtues (Loyalty, Subtle Thought, Gentle Gaze…) and characters borrowed from famous texts (Tristan and Iseut, Pyrame and Thisbé). The 14th century *Dame à la licorne et beau chevalier au lion* is a comparatively late example, of a romance mixed with lyrical poems, in a manuscript that was likely gifted to the princess Blanche de Navarre at the time of her wedding with the king of France. It is also an archetypal courtly story, in which a married young lady, accompanied by a unicorn, falls in love with a knight accompanied by a lion, and lives a story filled with tropes such as the rumors of death of her lover, the slanders of the jealous against the couple, etc.

The presence of the Song of Saint Alexis is seemingly a discrepancy in comparison to the rest of the list, yet it might be explained by the centrality in the tale of the marriage of Alexis, from which he flees, up to the end of the narrative, that finishes with the lamentations of his wife before his dead body (that creates a discordant echo to the courtly 'death from love').

On the other hand, the witnesses closest to the violence vector are chiefly epics (*chansons de geste*). For instance, *Aliscans* and the *Chevalerie Vivien*, that appear several times in the list, are centred around the eponymous battle in Aliscans between the Sarracen king Deramé of Cordoba and the Frank knight Vivien, who swore never to back down before the pagans, and endures an heroic death precisely because of his vow. It is interesting to notice the presence in this list, among more fictional texts, of the *Conquest of Jerusalem*, that draws on the events of the First Crusade (in particular, the siege of Jerusalem in 1099).

The nature of the documents closest to the love and violence vectors, when confronted to existing literary knowledge, confirms the quality of the document embedding, and the ability of our method to recognise the importance of courtly love or chivalric violence contents in the texts.

Table 4
The witnesses with the highest mean love (left) and violence (right) scores.

Work	Shelfm.	Date	*amer*	Work	Shelfm.	Date	*ferir*
Roman de la poire	fr. 12786	1281-1320	0.7	Anseïs de Carthage	fr. 368	1301-1313	0.54
Lai de l'Ombre	fr. 14971	1301-1400	0.66	Gaydon	fr. 15102	1201-1250	0.54
Lai de l'Ombre	fr. 837	1278-1280	0.65	Ogier le Danois	fr. 24403	1281-1300	0.54
Lai de l'Ombre	fr. 1593	1251-1300	0.65	Aliscans	fr. 774	1251-1300	0.54
Floris et Lyriopé	Ars. 5201	1300	0.64	Maugis d'Aigremont	fr. 766	1296-1305	0.54
Lai de l'Ombre	fr. 12603	1301-1320	0.63	Elie de Saint Gilles	fr. 25516	1251-1300	0.53
Cleomadés	fr. 24404	1281-1400	0.63	Aliscans	fr. 1448	1241-1260	0.53
Lai de l'Ombre	fr. 19152	1301-1320	0.62	Aliscans	fr. 368	1301-1313	0.53
Cleomadés	Ars. 3142	1281-1300	0.62	Chevalerie Vivien	fr. 368	1301-1313	0.53
Lai de l'Ombre	NAF 1104	1281-1300	0.62	Aliscans	fr. 1449	1241-1260	0.53
Amadas et Ydoine	fr. 375	1288	0.61	Garin de Monglane	fr. 24403	1281-1300	0.53
L'Escoufle	Ars. 6565	1276-1300	0.61	Bataille Loquifer	Ars. 6562	1225	0.52
Chanson de st Alexis	fr. 25408	1267	0.61	Anseïs de Carthage	fr. 12548	1201-1300	0.52
Eracle	fr. 1444	1281-1300	0.61	Folque de Candie	fr. 25518	1201-1300	0.52
Floire et Blancheflor	fr. 1447	1285-1295	0.60	Aspremont	NAF 10039	1241-1260	0.52
Lai de l'Ombre	fr. 1553	1284	0.60	Moniage Rainouart	fr. 1449	1241-1260	0.52
Yvain	fr. 12603	1301-1320	0.60	Roman d'Alexandre	fr. 12567	1301-1320	0.52
Berte aus grans piés	fr. 24404	1281-1400	0.60	Gaydon	fr. 860	1260-1290	0.52
Berte aus grans piés	Ars. 3142	1281-1300	0.59	Aliscans	Ars. 6562	1225	0.52
Dame à la licorne	fr. 12562	1349-1350	0.59	Chevalerie Vivien	fr. 1449	1241-1260	0.52
Cristal et Clarie	Ars. 3516	1267-1268	0.58	Roman d'Alexandre	fr. 368	1301-1313	0.52
Berte aus grans piés	fr. 12467	1281-1300	0.58	Gerbert de Metz	fr. 1443	1201-1300	0.51
Chanson de st Alexis	fr. 12471	1301-1305	0.58	Folque de Candie	fr. 774	1251-1300	0.51
Berte aus grans piés	fr. 1447	1285-1295	0.57	Auberi le Bourguignon	fr. 24368	1298	0.51
Charlemagne	fr. 778	1301-1325	0.57	Conqueste de Jerusalem	fr. 1621	1246-1255	0.51

3.3. Median scores variation in time

Examining the variation through time of the semantic contents of the documents, year by year, by plotting the average yearly document similarity of the samples with the vectors of the love and violence sets of keywords, seems to yield a strong increase of the presence of love until, roughly, the years 1330-1340, followed by a tendencial decrease, until the end of the Middle Ages, roughly coinciding with the Crisis of the Late Middle Ages (or the Medieval Great Depression), though not completely. A comparison with reconstructed economic data [9] shows that the important crises of the beginning of the 14th century, in particular the Great Famine of 1315–1317, that coincides with a very large drop in estimated GDP per capita, does not seem to affect the importance of love in fiction, though it might have contributed to the very significant drop in available manuscripts observed above (Figure 1). If there seems to have shared trends, up to a point, in long term variation of economic development and love score (increase in the 13th century, decrease during the Crisis of the Late Middle Ages), the comparison of the two curves do not match perfectly, and perhaps hints at a time lag of a couple of decades in the latter. This might hint at a form of cultural inertia, especially in a context where textual transmission (the copy and circulation of texts) is a lengthy process, and by no means as fluid as in latter periods. It is to be noted that some of the increases of the GDP per capita are not necessarily due to increase of GDP, but instead to sudden decreases in population, such as the one caused by the Black Death in 1347-1351.

Violence, on the other hand, seem to start its decrease earlier, around the middle of the thirteenth century. This could coincide with the slow loss of favour of the genre of epics

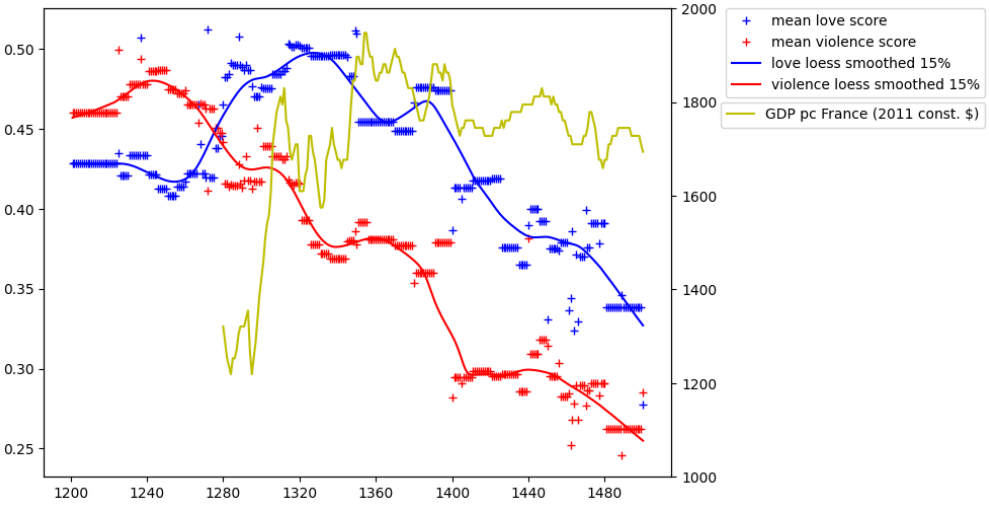


Figure 2: Diachronic variation of the mean passages score for the love and violence sets of keywords, in comparison to the variation of the GDP per capita (in 2011 constant dollars; data: [9])

(*chansons de geste*), victim of the competition of the more recent genre of *romans*, as well as the irruption in later *chansons de geste* of themes other than war: for instance individual adventure, love or wonder.

4. Discussion and future work

In building the corpus used for this study, we remain tributary of biases of the unequal preservation of documents through time, of large and small scale historical events, from the Great Plague to the ups and downs of the Royal Library, whose collections are the ancestor of those of the BnF that we used (cf. Figure 1). In addition, since manuscripts (especially those preserved) were expensive objects reserved to a certain elite, their contents cannot be claimed to represent the taste of society as a whole, but rather those of relatively wealthy and educated class (aristocratic or otherwise).

Yet, within the limits of these sources, we observe that we are able to build and query a semantic representation of the words and documents that exhibits many of the tropes of this literature, that researchers have studied through close reading. In particular, the semantic environment of the love word vector, both in terms of close words or documents, corroborates and sometimes enriches literary knowledge on the tropes of courtly love and the associated works. They align with several of de Rougemont's ideas about the importance of the lyric tradition, as well as the strong link between the themes of love, love induced suffering and death.

The variation in time of the importance of love and violence shows initially opposite trends, that, after c. 1340, seem to align more closely. In terms of literary history, this could correspond to the traditional epic *chansons de geste* focused on collective war against sarracens and feudal conflicts slowly going out of fashion, and progressively aligning their content with the more

modern genre of the *roman*, including individual adventures and love stories. Once epics have merged with chivalric romances, both seem to behave in similar ways through time.

Finally, the variation in time of the mean importance of love in the fictions seem to show a phase of expansion until the early 14th century, when it then knows a downward trend during the period of the Crisis of the Late Middle Ages (with a time lag of roughly 20 years, the decline in love starting around 1330-1340, while the Great Famine of 1315-1322 traditionally marks the beginning of the Crisis). Further research is needed to explore this issue in greater depth, and test the correlation with economic development as well as other factors.

Data and materials availability

Data and scripts used for topic modelling are available on a Zenodo repository: 10.5281/zenodo.10011791. The CMFER is also available on Github, https://github.com/Jean-Baptiste-Camps/CMFER.

References

[1] C. Seignobos, L'Amour est-il une invention moderne ?, Le Quotidien (1925).

[2] D. de Rougemont, L'Amour et l'Occident, republ. online in Rougemont 2.0 (genève) ed., Paris, 1939. URL: https://www.unige.ch/rougemont/livres/ddr1939ao.

[3] A. Corbellari, Retour sur l'amour courtois, Cahiers de recherches médiévales - Journal of medieval studies (2009) 375–385. doi:10.4000/crm.11542, number: 17 Publisher: Classiques Garnier.

[4] N. Baumard, The Ancient Literary Fictions Values Survey, OSF (2021). URL: https://osf.io/mvybs, publisher: Open Science Framework.

[5] C. Morris, The discovery of the individual, 1050-1200, volume 5, University of Toronto Press, Toronto, 1987.

[6] N. Baumard, E. Huillery, A. Hyafil, L. Safra, The cultural evolution of love in literary history, Nature Human Behaviour 6 (2022) 506–522. doi:10.1038/s41562-022-01292-z, number: 4 Publisher: Nature Publishing Group.

[7] G. Duby, Mâle Moyen Âge: de l'amour et autres essais, Flammarion, Paris, France, 1987. ISSN: 0768-1011.

[8] L. Ridolfi, The French economy in the longue durée: A study on real wages, working days and economic performance from Louis IX to the Revolution (1250–1789), Ph.D. thesis, IMT School for Advanced Studies, Lucca, 2016. URL: http://e-theses.imtlucca.it/211/1/Ridolfi_phdthesis.pdf.

[9] J. Bolt, J. L. Van Zanden, Maddison style estimates of the evolution of the world economy. a new 2020 update, Maddison-Project Working Paper WP-15, University of Groningen (2020).

[10] M. Kestemont, F. Karsdorp, E. de Bruijn, M. Driscoll, K. A. Kapitan, P. Ó Macháin, D. Sawyer, R. Sleiderink, A. Chao, Forgotten books: The application of unseen species models to the survival of culture, Science 375 (2022) 765–769. doi:10.1126/science.abl7655, publisher: American Association for the Advancement of Science.

[11] J.-B. Camps, J. Randon-Furling, Lost Manuscripts and Extinct Texts: A Dynamic Model of Cultural Transmission, in: Proceedings of the Computational Humanities Research Conference 2022 Antwerp, Belgium, December 12-14, 2022, CEUR Workshop Proceedings, 2022, pp. 198–214. URL: https://ceur-ws.org/Vol-3290/long_paper3261.pdf.

[12] L. Ing, L'obsolescence lexicale en français médiéval: Philologie et linguistique computationnelles sur le Lancelot en prose, Phd thesis, Université Paris Sciences et Lettres, 2023. URL: https://www.theses.fr/s221114.

[13] J.-B. Camps, S. Gabay, G. F. Riva, Open stemmata: A digital collection of textual genealogies, in: EADH2021: Interdisciplinary Perspectives on Data, 2nd International Conference of the European Association for Digital Humanities, 2021. URL: https://halshs.archives-ouvertes.fr/halshs-03260086.

[14] Bibliothèque nationale de France, catalogue BnF Archives et manuscrits, 2023. URL: https://archivesetmanuscrits.bnf.fr/.

[15] IRHT, Jonas: Répertoire des textes et des manuscrits médiévaux d'oc et d'oïl, 2023. URL: http://jonas.irht.cnrs.fr/.

[16] L. Brun (Ed.), Arlima - Archives de littérature du Moyen Âge, 2005. URL: https://www.arlima.net/.

[17] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[18] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.

[19] B. Karas, S. Qu, Y. Xu, Q. Zhu, Experiments with lda and top2vec for embedded topic discovery on social media data—a case study of cystic fibrosis, Frontiers in Artificial Intelligence 5 (2022) 948313.

[20] J. H. Lau, T. Baldwin, An empirical evaluation of doc2vec with practical insights into document embedding generation, in: Proceedings of the 1st Workshop on Representation Learning for NLP, 2016, pp. 78–86.

[21] S. A. Curiskis, B. Drake, T. R. Osborn, P. J. Kennedy, An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit, Information Processing & Management 57 (2020) 102034.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).

[23] C. Ruby, Thibaut, in: Dictionnaire des lettres françaises: Le Moyen Âge, 1992, pp. 1422–1423.

[24] T. Clérice, You actually look twice at it (yaltai): using an object detection approach instead of region segmentation within the kraken engine, arXiv preprint arXiv:2207.11230 (2022).

[25] A. Pinche, K. Christensen, S. Gabay, Between automatic and manual encoding, in: TEI 2022 conference: Text as data, 2022.

[26] S. Gabay, J.-B. Camps, A. Pinche, C. Jahan, Segmonto: common vocabulary and practices for analysing the layout of manuscripts (and more), in: 1st International Workshop on Computational Paleography (IWCP@ ICDAR 2021), 2021.

[27] B. Kiessling, Kraken - an Universal Text Recognizer for the Humanities, in: Digital Humanities Conference 2019, Complexities, Utrecht (DH2019), 2019. URL: https://web.

archive.org/web/20210719115330/https://dev.clariah.nl/files/dh2019/boa/0673.html.
[28] T. Clérice, A. Pinche, M. Vlachou-Efstathiou, Generic CREMMA Model for Medieval Manuscripts (Latin and Old French), 8-15th century (2023). doi:10.5281/zenodo.7631619, publisher: Zenodo.
[29] A. Pinche (Ed.), Guide de transcription pour les manuscrits du Xe au XVe siècle, Paris, 2022. URL: https://hal.science/hal-03697382/.
[30] T. Clérice, Ground-truth Free Evaluation of HTR on Old French and Latin Medieval Literary Manuscripts, in: F. Karsdorp, A. Lassche, K. Nielbo (Eds.), Proceedings of the Computational Humanities Research Conference 2022 Antwerp, Belgium, December 12-14, 2022, volume 1613 of *CEUR Workshop Proceedings*, Antwerp, 2022, pp. 1–24. URL: https://ceur-ws.org/Vol-3290/long_paper2081.pdf.
[31] D. Angelov, Top2vec: Distributed representations of topics, arXiv preprint arXiv:2008.09470 (2020).
[32] P. Ma, Q. Zeng-Treitler, S. J. Nelson, Use of two topic modeling methods to investigate covid vaccine hesitancy, in: Int. Conf. ICT Soc. Hum. Beings, volume 384, 2021, pp. 221–226.
[33] R. Egger, J. Yu, A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts, Frontiers in sociology 7 (2022) 886498.
[34] J. Van Zundert, M. Koolen, J. Neugarten, P. Boot, W. Van Hage, O. Mussmann, What Do We Talk About When We Talk About Topic?, in: Proceedings of the Computational Humanities Research Conference 2022 Antwerp, Belgium, December 12-14, 2022, CEUR Workshop Proceedings, Antwerp, 2022, pp. 398–410. URL: https://ceur-ws.org/Vol-3290/short_paper5533.pdf.
[35] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).
[36] L. Grobol, M. Regnault, P. O. Suarez, B. Sagot, L. Romary, B. Crabbé, Bertrade: Using contextual embeddings to parse old french, in: 13th Language Resources and Evaluation Conference, 2022.

A. Acquisition workflow and evaluation of the corpus

A.1. Workflow

The workflow for text acquisition is depicted in fig. 3.

Manuscripts images are harvested using the International Image Interoperability Framework (IIIF), based on their manifest, then processed through layout analysis, using YALTAi [24] object detection approach, and the Gallicorpora model [25], using SegmOnto ontology for the semantic typing of zones [26], in combination with a Kraken [27] model for the identification of lines. The resulting ALTO (Analyzed Layout and Text Object)/page images pairs are then passed to handwritten text recognition, using the deep learning approach of the Kraken software, and the CREMMA Medieval Generic model [28]. This model produces a version of the text that encodes abbreviations as such, and follows the graphematic conventions recently elaborated at the École des chartes, in a seminar led by A. Pinche, J.-B. Camps and F. Duval

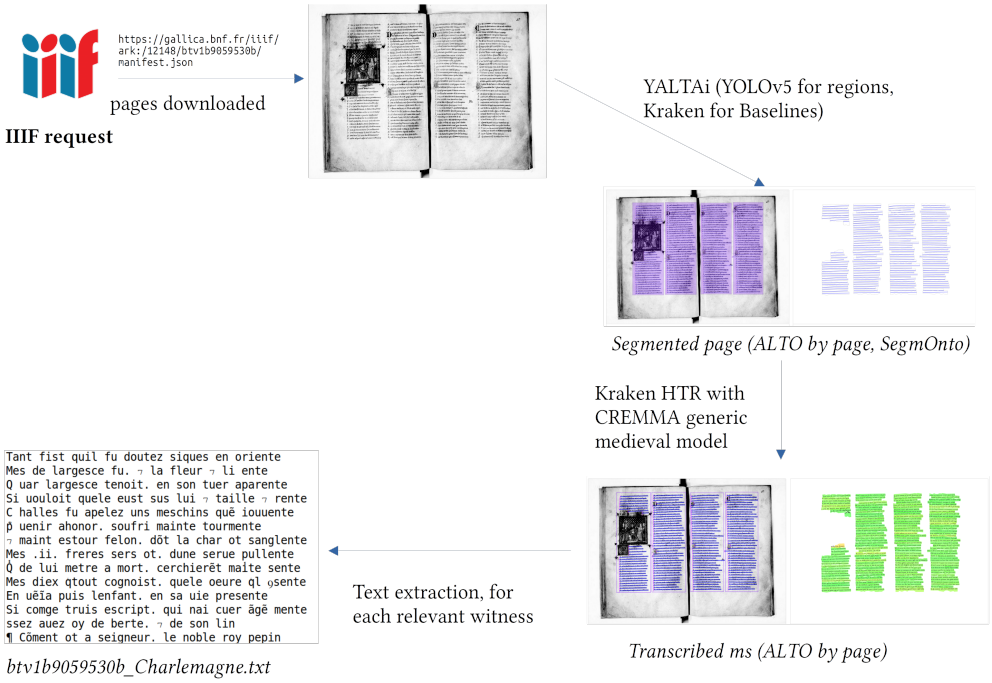


Figure 3: Workflow for the production of the corpus.

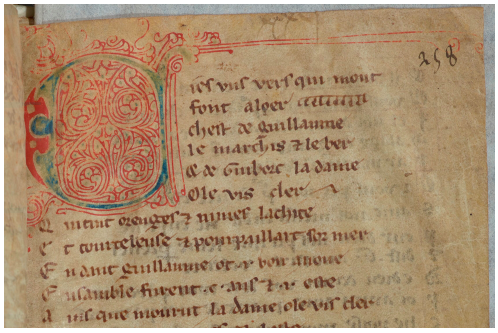
[29].

The resulting ALTO files (one per page) are then processed through a dedicated script, to create a single raw text file per witness (i.e., an instance of a given work in a given manuscript), with the relevant metadata in an accompanying tsv format.

A.2. Quality evaluation

We follow the approach recently described by Clérice [30], for ground-truth free evaluation of handwritten text recognition (HTR) of Old French. This approach is based on natural language processing, and aims to evaluate the apparent linguistic consistency of a text, rather than its match with the original line image. It takes the evaluation as a classification task, were a model is trained to classify transcribed lines in categories, that are supposed to approximate a level of character error rate: Good ([0, 10)%), Acceptable ([10, 25)%), Bad ([25, 50)%), and Very Bad ($\geq$ 50%). For this, it uses a model based on a an embedding-sentence encoder-linear classifier structure. It produces as an output a classification of each line in each of the aforementioned categories (fig. 4). We reuse the model provided by Clérice with the original paper.

To provide an estimate for each textual witness, we count the total number of lines in each categories, and compute a ratio, both for each category, but also for good and acceptable vs bad and very bad (fig. 5). median ratio of good lines is 65% and the median ratio of good+acceptable lines is 94% (min: 9%; 1st quartile: 89%, 3rd quartile: 97%; max: 100%). Typical examples of



<lb cert="Acceptable"/>ies uns uers qui mont [CER: 4.7%]
<lb cert="Bad"/>font aloer aalucid [CER: 70%]
<lb cert="Acceptable"/>chest de quillaume [CER: 5.6%]
<lb cert="Good"/>le marchis ⁊ le ber [CER: 0%]
<lb cert="Bad"/>&c' de guborc la dame [CER: 10%]
<lb cert="Bad"/>Pole uis cler [CER: 10%]
<lb cert="Acceptable"/>Q uitint orenges ⁊ nimes lachite [CER: 0%]
<lb cert="Acceptable"/>E t couiteleuse ⁊ poinpaillait sor mer [CER: 13%]
<lb cert="Acceptable"/>E n dant guillaume ot .x. bon anoue [CER: 5.7%]
<lb cert="Acceptable"/>E usamble furent .c. ans ⁊ .i. este [CER: 2.8%]
<lb cert="Acceptable"/>A ins que mourut la dame ole uis cler [CER: 0%]

Figure 4: The beginning of *Moniage Guillaume* in Arsenal, 6562, fol. 258r, and the HTR output, with the classification performed, and the actual character error rate. Here, qualitative inspection shows that the classification actually tend to be more pessimistic than the actual error level (all misclassifications are down one category).

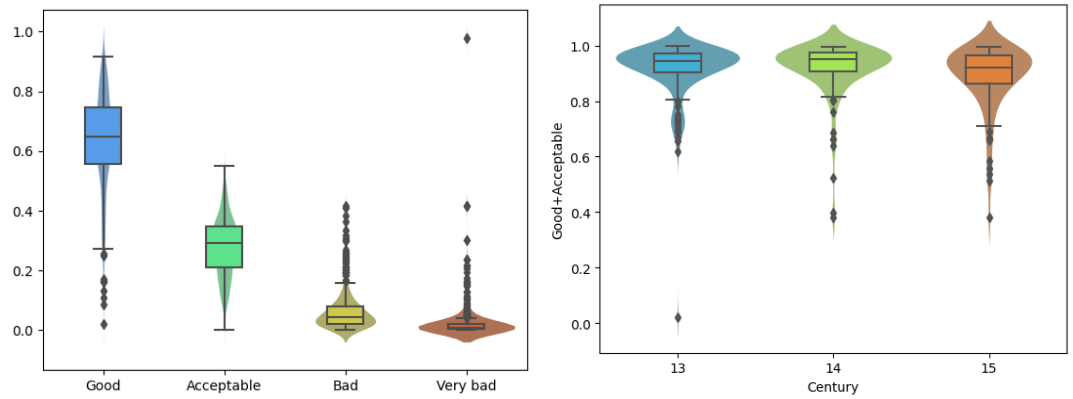


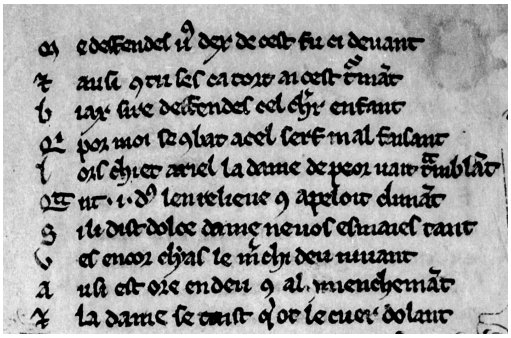
Figure 5: Box and violin-plots for the textual witnesses of the corpus, showing the ratios of lines categories assigned by the HTR quality evaluation.

results for maximum, median, first quartile and minimum values of good+acceptable lines are given in Appendix B.

Distribution of quality estimations by century shows that our model shows comparable levels of quality for the 13th and 14th century, with most manuscript above 80% of good and average lines, and a few outliers below. On the other, there is a decrease in quality for the 15th century, with also a less compact distribution. This can be explained by the significant number of 15th century manuscripts written in cursive scripts, with often less formal execution, that differ significantly from the Gothic *Textualis* that otherwise dominates the corpus.

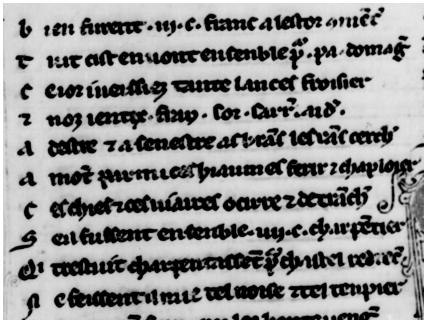
Outliers with a large number of bad or very bad lines exists, and they were removed from the corpus before further analysis. The threshold was set at 1.5 interquartile range below the 1st quartile (ratio of good+average lines ≥ 0.78), resulting in 370 texts selected for further analysis (Table 1).

B. Example of processing results for the different quality levels



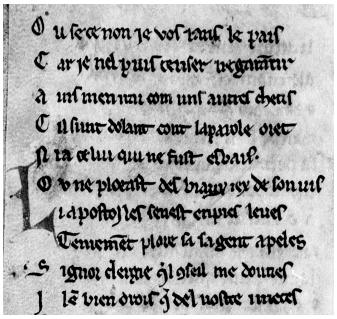
M edeffendes u^{9} dex de cest fu ci deuant
⁊ ausi 9tu ses ca **c**ort ai cest c̊mãt
b iax sire deffendes cel ch̓r enfant
Q̓ por moi se 9ba**r** acel serf mal faisant
l ors chiet ariel la dame de peor uait d̊mblãt
9tt ut .i. d^{9} len relieue 9 apeloit climãt
s ili dist dol**o**e dame neuos esmaies tant
U es encor elyas le m̓chi deu **n**iuant
A usi est ore endeu 9 al **mi**enchemãt
⁊ la dame se taist q̓ ot le cuer dolant

Figure 6: BnF, fr. 1621 (c. 1250), fol. 2r, *Chevalier au cygne* – **Ratio of good+acceptable lines, 98.86%** (errors in bold), between third quartile and maximum.



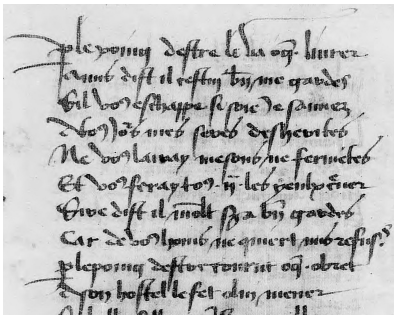
b ien furen**i**t .iii.**e**. franc a **i**estor 9mẽc̓
T uit cistenuont ensenble p̃ .pa. domag̓
c eiorin**e**issie **d**an**c**e lances froisier
⁊ noz ien**g**e . **s**ra**p**. sor. sarr̓. aid'.
a destre ⁊ a sen**c**stre a**c**rrã**c** **i**esrãs cer**a**
a mot parmices hiaumes **s**e**u**r ⁊ chaploier
c eschie**t**i**c**e**cus**aires ocirre ⁊ de **c**rãch̃^{e}
Se**u** **lui**rent en**c**enble .iiii.**e**. charpẽtier
Qi trestuit **qui**pentata**ẽ** **f**chist**e**i re**ric**ẽ
Ne **s**eissent**a** mie tel**u**oise ⁊ te**i**tenpier

Figure 7: BnF, NAF 6298 (13th century), fol. 4v, *Siege de Barbastre* – **Ratio of good+acceptable lines, 94.28%** (errors in bold), median individual.



O u se**s** non ie uos rans **k**e pais
C ar ie nel puis **au**ser **u**egarãtir
a **un**s men **n**ai com uns autres che**a**s
C il sunt dolant cont lapa**i**ole o**u**t
N **r**a celui qui ne fust **a**bais.
O une plorast de**l** biaux iex de sonuis
i aposto**s** les senest enpies leues
Cememẽt plore sa sagent apeles
S ignor clergie q̓**l** 9seil me dou**u**es
I lẽ bien **o**rois q̓ del uostre **uue**ces

Figure 8: BnF, fr. 1622 (13th 3/4), fol. 2v, *Garin le Loherain* – **Ratio of good+acceptable lines, 91.62%** (errors in bold), between first quartile and median.



ple**xc**in destre **L**a **cq̃** liurer
Iau**n**s dist il cest**qi** bñ me gardes
uil uo^{9} eschap**t**e si **s**e**i**e Ie sa**i**mez
duo^{9} Iõs mes seres desheri**b**es
Ne uo**i**lairay mes**t**u**s** ne feri**e**les
Et uo^{9} feray to^{9} .ĩ. les yeulx c̊er
Sire dist il molt s**g**a bñ gardes
Car de uoi**h**a**i**s ne quie**c**t **an**s refus**i**
plepoing des**tot** **c**eue**n**t oq̃. obre**t**
son h**e**stelle fet **din** mener

Figure 9: BnF, fr. 1583 (15th c.), fol. 1v, *Ogier le Danois* in decasyllabic verses – **Ratio of good+acceptable lines, 38.03%** (errors in bold), outlier close to the minimum level (5th text, starting from the lowest).

C. Composition of the love and violence vectors

C.1. Love

All forms of verb ‘aimer’ (to love) and noun ‘amour’ (love) that were found were used. They are the following (forms with HTR errors are marked with an asterisk):

> aime, aimme, ama, amast, ame, amee, amer, amerai, ameroit, ames, *amo, amoit, amor, amors, amour, amours, *anier, *anne, *camoi, damor, damors, damour, damours, desamour, iaim, iaime, iamoie, laim, laime, lamasse, lamerai, lamoit, lamor, lamour, *laune, maime, mamast, *mor, *mour, *mours, naim, naime, namerai, quamours, samor, samors, samour, *sanie, taime.

C.2. Violence

All forms of verb ‘ferir’ (to hit) that were found were used. They are the following (forms with HTR errors are marked with an asterisk):

> anfiert, efiert, enfiert, feri, ferir, feru, ferus, fiert, leferi, referi, refiert, *uaferir

D. Topic modelling

D.1. top2vec model

In addition to word and document embeddings, we investigated the texts using top2vec [31]. Recent studies have shown top2vec to yield qualitatively better results, and more coherent and human-readable topics than other topic modelling methods, such as the classic LDA [32, 33, 19]. It has been already used in large scale topic modelling of literary corpora [34].

In addition, top2vec automatically finds the relevant number of topics, which will facilitate the handling of this large corpus by relieving us of doing long and computationally intensive benchmarks of arbitrary number of topics.

top2vec was trained reusing the doc2vec model described in the main text, with otherwise default hyperparameters. The ϵ parameter for the dbscan clustering of topics was set to 0.1 in cosine distance (i.e., topic vectors with a smaller cosine distance will be merged).

In the lack of benchmarks dedicated to variation-rich historical corpora similar to ours, we still conduced some degree of experimentation on the variation of these parameters (e.g., using longer n-gram vocabulary, adjusting ϵ, using top2vec ‘fast-learn’, ‘deep-learn’ alternatives, etc.), but not in a systematic fashion, due to the long training time for each model (up to 10 hours). Experiments resulted in apparently lower quality topics, with either a excessively small number of topics (e.g., 5 topics) or less significant topics with a predominance of function words.

The training with the chosen parameters yielded 276 topics.

We chose top2vec over BERTopic [35], due to the unavailability of a pretrained BERT model compatible with the specifics of our data, in terms of language and writing conventions, e.g., abbreviations (see next subsection) .

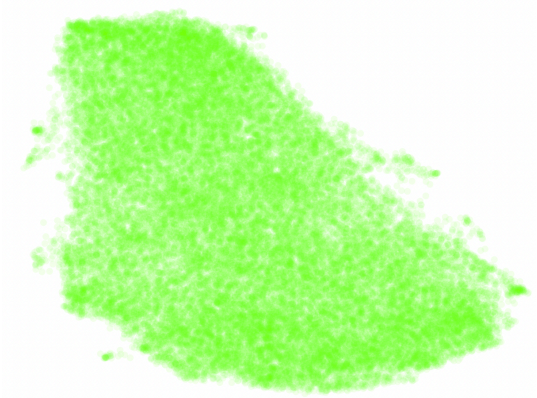


Figure 10: Semantic map (umap) of the corpus embeddings with paraphrase-multilingual-MiniLM-L12-v2 displays a lack of distinctive clusters

D.2. Experiments with BERTopic

BERTopic was another option that shared many of the strength of top2vec and performs especially "well on most aspects of the topic modeling domain"[33, 12]. BERTopic can run on any pretrained BERT model but is commonly associated with a multilingual pre-trained embedding model trained on Reddit and StackExchange, paraphrase-multilingual-MiniLM-L12-v2. Preliminary tests showed that the model is affected by *historical drift*, due to the increasing distance between older version of French written languages and the contemporary standard: BERTopic did run correctly on a set of 17th century French novels and to a lesser extent on a large sample of 15th century texts from our corpus. Before the 15th century, the results were totally inconclusive with one topic containing nearly all the corpus. The semantic map in fig. 10 suggests that sentence embeddings have deteriorated to such an extent that it is no longer possible to recover regular clusters of topics.

We plan to pursue experiments with this method, to be able to compare its results with those presented here, once a pre-trained embedding model fitting our data is made available or is trained by us. Indeed, if there exists an Old French Bert, *BERTrade* [36], but it is based on a corpus significantly smaller that the data we gathered (10 million words, as opposed to 40), and more importantly uses text editions with an higher level of normalisation (abbreviations expanded, in particular).

This call for a methodological remark: the rapid development of masked language models following BERT has created a new range of issues for historical studies. Most models are trained on very recent corpus and data. They are unlikely to cover past linguistic forms and writing and using pre-trained models alone, it would not yet have been possible to conduct this study.

D.3. Results of top2vec

The six topics which scored higher for the semantic similarity with the vector of 'love' word forms are shown in fig. 11. Interestingly, the two highest scoring topics both relate to the lyrical register of the complaint (*plainte*) for the pains and injuries caused by love and the act

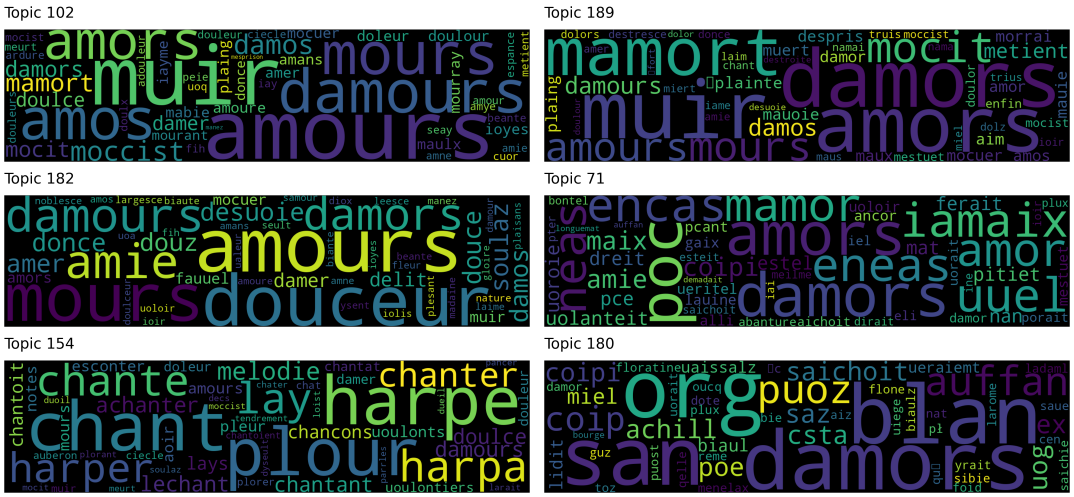


Figure 11: The topics most correlated to the love keywords. The first two topics (love score for both of 0.33) are related to the **complaint for the pains of love and for dying of love**; the third topic (0.31) is dedicated to the **sweetness, comfort and pleasure** (*douceur*, *solas* and *delit*) of love; the fourth topic (0.29) is mostly related to the **story of *Dido and Aeneas***, as found in the *Roman d'Eneas*; the fifth (0.20) topic deals mostly with the songs of love, and particularly the story of the *Lais* in the *Tristan en prose*, in particular **the *Deadly lai***; finally, the sixth topic (0.19) is related to the story of ***Blancandin et l'Orgueilleuse d'Amour***.

of (metaphorically) dying of love / being killed by love as they are found abundantly in our corpus inside the *Tristan en prose* and its many lyrical poetry inserts. The death is again found in the fifth topic, that concerns the songs of love in general, but also very particularly the *Lais* of the *Tristan en prose*, especially the *lay mortel* (*Deadly lai*), the last love song sung just before dying of that same sentiment.

In apparent strong contrast, the third topic appear dedicated to the pleasures of love, its sweetness and the comfort it brings, through it can be closely nested with the previous one, as is demonstrated in one of the highest scoring passages, taken from a lyrical (and possibly parodic) part of the *Romans de Fauvel*, of which we give here an abstract with minor corrections to the HTR:

> Q ue ie muir par tres bien amer
> E n ce que urai martir serai
> D ame en mourant me reconforte
> *[My lady please remember that] I am dying because of loving very well, that I will be a true martyr of love, lady, this brings me comfort while I die.*

Other passages were this topic is most represented are found in a variety of sources, from *Amadas et Ydoine* to the *Roman des Sept Sages de Rome* (*Seven Wise masters*) and its continuations.

Finally, the fourth and sixth topics are related to specific works, the *Romans d'Eneas* and a somewhat less known and perhaps overlooked work, *Blancandin et l'Orgueilleuse d'Amour*..

Figure 12: The topics most correlated to the violence (*ferir*) keywords. Though they are all related to battle scenes, the first one is about **slicing one's ennemy and his armour**, while the second one is more concerned with **brandishing one's spear and piercing the ennemy shield**, the third one by **throwing him off his horse and to the ground**, injured, knocked out or dead (optionally after skewering him); the fourth relates to the fight inside a specific work, *Guiron le courtois*, the fifth with **swordfighting**, and the sixth with **spearfighting**.

Topics most related to violence concern, in an unsurprising manner, descriptions of battles and fights between a knight and his enemies. They are relatively straightforward to interpret and concern different but related aspects of knightly sword- and spear-fighting (fig. 12).