

Modelling the impact of censorship on early modern literary texts

Enrica Zanin^{1,*†}, Demian Battaglia^{2,†}

¹University of Strasbourg, 14, rue Descartes, 67000 Strasbourg, France

²CNRS, LNCA UMR 7364, FunSy team, 12 rue Goethe, 67000 Strasbourg, France

Abstract

How can we quantify the impact of censorship on the style and content of literary texts? This article aims to answer this question by considering one of the work that has been most widely censored and rewritten, that is, Boccaccio's *Decameron*. At the time of the rise of censorship in the sixteenth century, the *Decameron* was banned by the roman index and rewritten in three censored versions. By comparing these different versions in an unsupervised manner, using latent semantic analysis, the aim is to understand the directions in which censorship caused these texts to evolve, in order to understand whether the major changes in narrative practice that occurred at the end of the 16th century were partly due to, or at least supported by, censorship.

Keywords

Decameron, early modern Censorship, latent semantic analysis

1. Introduction

At the end of the 16th century, censorship structures were gradually put in place in Europe: the Church of Rome and the French, Spanish and British monarchies created, at different times and in different ways, structures to control the production and circulation of books (see Fragnito 2019, Frajese 2008, Clegg 1997, Hadfield 2001, Landi 2011, Minois 1995, Pinto Crespo 1983). But the censors did not limit themselves to influencing the quantity of books in circulation: they also affected their content, form and development over time. Unfortunately, the impact of their actions is difficult to detect because censorship documents, the stages of negotiation between censors and authors, and the various provisional versions of literary texts are generally erased. In most cases, only the final, published version of the work remains. How can the influence of censorship be reconstructed?

Literary critics agree that literary forms changed from the end of the sixteenth century onwards, but generally explain this by the change in readers' "taste" (Plazenet 1997). However, it is highly likely that, if readers' tastes changed, it was because cultural prescribers (and censors in particular) encouraged and supported this change.

To measure the impact of censorship on changes in literary forms at the end of the sixteenth century, we therefore need to compare the censors' indications (which appear clearly in the Church indexes and censorship manuals) with the changes made to censored texts, to see whether these changes are in line with the evolution of literary forms at the time. If we were able to carry out this comparison, the question would still be open as to whether censorship was the cause of these changes or whether it merely encouraged them. In either case, however, it would be possible to measure its impact.

To begin this study, we looked at one of the genres most heavily criticized by censors, the short story, whose immorality and impiety were deplored by moralists and censors alike. At the center of the controversy was the collection of short stories that founded the genre: Boccaccio's *Decameron* (written in 1350), banned by the Roman Index in 1559 and republished in three censored versions: The first by a group of Florentine academics (called *Deputati*), led by Vincenzo Borghini and supervised by the Master of the Sacred Palace, Thomas Manrique (printed in Florence by Giunti in 1573), the second by the Florentine writer, protected by the Archiduc of Florence, Alessandro Salviati (published by Giunti

8th International Workshop on Computational Models of Narrative

*Corresponding author.

†These authors contributed equally.



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

in 1582), the third by Luigi Groto, a blind scholar and playwright (published in Venice in 1588). Also preserved is the correspondence between the censors and the Florentine academicians who prepared the first censored edition (Deputati, 1573). These exchanges, along with the four versions of the text (the original and the three censored versions), can serve as a starting point for understanding the impact of the censors on literary practice and its development. What poetic choices were made in the three censored versions? How do these choices reflect (or not) changes in readers' tastes? To answer these questions, the method chosen is to compare reading impressions (based on an in-depth knowledge of the texts and context) with the unsupervised results of a quantitative study conducted in collaboration with interdisciplinary data science algorithms specialists.

2. Textual signs of change

An intuitive analysis of the versions of the *Decameron* reveals some interesting signs of change. On the one hand, the three censored versions seem to follow the rules of rewriting, set out explicitly in the *instructio*¹ that accompanies the Index of Rome, albeit through different rewriting strategies. On the other hand, the narrative experimentation at work in the censored versions of the *Decameron* reflects the literary trend of the late sixteenth century.

2.1. Characters change: secularization and racialization

The most obvious changes concern certain words that disappear from the novellas. The “priests”, “monks”, “nuns”, and other clerics who populated the comic universe of the short story are replaced by lay figures. The *Instructio* accompanying the 1596 Index stipulated that all references to religious orders and rites were to be removed². The censorship of the *Decameron* bears witness to this: Friar Cipolla, who uses relics to deceive the faithful (*Decameron* VI, 10), is reduced to a layman in Borghini's version of the *Decameron*, which deliberately omits the word “frate”; for Salviati, he is a brigand (*ribaldo*) disguised as a monk; for Groto, Friar Cipolla is a peddler (*ciurmatore*)³.

In the copy of the *Decameron* directly annotated by Thomas Manrique, the Master of the Sacred Palace in charge of censoring the text, all words related to the Christian religion are highlighted in red ink. In Ferondo's novella (III, 8), “abbate”, “monaco”, “confessarsi”, “purgatorio”, and “santo” are highlighted in red, as are “chostro”, “sepoltura”, “penitenza”, “seppellito”, and “risuscitato”⁴, which disappeared in the first censored edition of the *Decameron* (Deputati 1573)⁵. The disappearance of religious characters leads to a secularization of the narrative not only in the *Decameron* but also in later collections of novellas: the monk and the nun, who were essential characters in the short story genre, disappear and are sometimes replaced by Jews or priests disguised as Jews, as in a short story by Giraldis Cinzio (*Ecatommiti*, I, 4). The secularization of the characters is sometimes accompanied by their racialization: in the context of the Counter-Reformation, the deviant characters are no longer monks and nuns, but Jews, Jewish converts, Gypsies, or Moors⁶.

2.2. Plot change: from ambiguity to poetic justice

Narrative must be unambiguous. The *Instructio* states that all ambiguous terms, whose meaning is uncertain or plural (*verba dubia et ambigua*)⁷, are to be erased. Not only the words, but also the virtue

¹These include ambiguous and equivocal words, statements with a pagan meaning, statements contrary to morality, degrading jokes, lascivious and obscene images. See Bujanda 1994, 348-351.

²Bujanda 1994, 350 in latin 927.

³Novella VI, 10, in Boccaccio, *Decameron*, ed. « deputati fiorentini », 1573, p. 339; Boccaccio, *Decameron*, ed. Lionardo Salviati, , 1582, p. 341; Boccaccio, *Decameron*, ed. Luigi Groto, 1588, p. 153.

⁴“abbot”, “monk”, “confess”, “purgatory”, “holy” “cloister”, “burial”, “penance”, “buried”, and “resurrected”.

⁵Giovanni Boccaccio, *Decameron*, Venezia, Aldo Romano and Andrea Asolano, 1522, call number : RARI.22.A.4.1, fol. 90v-101r. See Carrai Madricardo 1989, 225-247.

⁶See the appearance of Jew, Moors, Gipsy and convert Jew in the novellas by Cervantes and Solórzano in Spain, Belleforest and Camus in France.

⁷Bujanda 1994, 926.

or wickedness of the characters must be clear and unambiguous. This is probably why the master of the sacred palace, in charge of the revision of the *Decameron*, Thomas Manrique, asked the *Deputati* to remove any formula that might cast doubt on the good or bad qualities of the characters, so that in each story the good would deserve a reward and the bad would deserve a punishment⁸. The censorship of the *Decameron* reflects this change. If monna Filippa (VI, 7), who is accused of adultery, summoned before the *podestà* and found guilty, does not receive the punishment that her guilt demands but, on the contrary, returns home “in glory”, it is a different story in the censored versions. For Salviati, monna Filippa is not acquitted: she is convicted, but her sentence is reduced (“alla donna ... fu alleggerita la pena”, p. 336). For Groto, it is the husband who is punished for not controlling her, and Filippa must give up her adventures and live chastely with him (“et debbe poi viveri col marito in santa pace”, p. 151).

More generally, a cursory reading of the novellas published since the end of the sixteenth century seems to show that the plot tends to follow the logic of poetic justice. Poissenot makes it clear that all ambiguity is excluded from his *histoires tragiques* (1586)⁹; Robert Smythe explains that his novellas are “lively pattern” that shows, as in a “table”, the rise of virtuous souls and the fall of vicious ones¹⁰. Poetic justice becomes a common strategy for moralizing novellas, because it clearly identifies vice and shows why it should be avoided, and recommends virtue, which readers must follow in order to be rewarded.

The rise of tragedy, which became one of the most prominent genres after the reception of Aristotle’s *Poetics* and the translation of the ancient tragedies, contributed to the development of a “tragic” plot that tended to follow the logic of poetic justice, as the characters who had committed a “great fault” were punished, as seen in certain tragedies (see Rymer¹¹) and in many *histoires tragiques*¹².

2.3. Style change: from licentiousness to idealization

The censors demanded that all lascivious language and obscene terms be eliminated (“lasciva, quae bonos mores corrumpere possunt”¹³). For this reason, explicit terms, suggestive descriptions, and bawdy scenes were removed from the novellas, as the rewritten versions of the *Decameron* by Borghini, Salviati and Groto show¹⁴. The authors of the time did the same: while they developed the analysis of amorous feelings, the description of bodies disappeared. While Bandello, in one of his novellas (IV, 25), carefully described the slow undressing of a lady, this description disappears in a later version by Bargagli, who ends the scene with a skillful fade to black¹⁵.

To soften the immorality of the plot, the authors move it to faraway lands that ignore the dictates of Christian morality. The story is thus taken out of its ideological framework, removed in time and space, and ends up resembling an oriental or fairy tale. While in Boccaccio’s version Alatiel, kidnapped by pirates, experiences all the pleasures of the flesh and later presents herself as chaste and virginal to her future husband (*Decameron* II, 7), in later versions of the same story the lascivious young woman is sent to a convent (in Giraldi’s *Ecatommiti*, I, 10) or manages to preserve her virtue in the midst of savage pirates who lust after her (in Ascanio de Mori, *Novelle* 1, 1585).

⁸Letter of Thomas Manrique to deputati, 15.3.1572, in Chiecchi 1992, 151.

⁹“Les vices y sont blamez, on y loue la vertu, on appelle bois ce qui est bois, pomme ce qui est pomme, et n’y peut-on licitement desguiser ou masquer d’une fardée vertu ce qui n’est ni vertueux ni louable”, Poissenot, 1996, “prologue de l’auteur”, 49.

¹⁰“wherein be lyuely patternes set downe & expressed, testifying what gaynes is gayned by voluptuous lyfe, and what guerdon ensueth sensuall appetite: wherein as in a Table, is forcibly delineated, how some for vertue, haue rysen to honour, fame, dignitie, and renowne, and some for vyce haue bene throwne downe, and buried in the pit of dishonour, shame, reproch, and eternall infamy. So that the worthy aduancement of the one, maye be a sharp spurre to styrrer vp to goodnesse, and the lothsome beastlynnesse of the other, a brydle to reine our appetytes from these inconueniences”, Smythe, 1577, “T. N. to the curteous, friendlye, and indifferent Reader”, not paginated.

¹¹Thomas Rymer is the first to use the terms “poetic justice” to define the plot of a well-composed tragedy in *The Tragedies of the Last Age*, 1678.

¹²See the “tragicall histories” by Boaistuau, Belleforest Habanc, Poissenot in France, by Fenton and Smythe in Britain, by Millis Godínez and Maria de Zayas in Spain. See Pietrzak 2006, 8 ; Crémone 2019, 54-59.

¹³Bujanda 1994, p. 927.

¹⁴See, for example, their rewriting of the story of Alibech, *Decameron* III, 10.

¹⁵Bargagli, 1989, III, 1, 475.

From the end of the sixteenth century, the novellas tend to follow an *art de l'éloignement*. The stories are often set in exotic lands, the characters are idealized, their adventures moralized: the novella is influenced by the sentimental novel that have been raging in Europe since Amyot's translation of Heliodorus' *Etyopics*¹⁶. However, as Annabel Patterson has pointed out¹⁷, the success of the sentimental novel and of the pastoral is not only due to a change in readers' tastes, but also to the censors' demand that plots be moralized.

3. Towards a quantitative model of literary change

The three signs of change that I have listed above are the fruit of reading the collections of short stories published in Italy, France, and Spain from the end of the sixteenth century to the end of the seventeenth century, of analyzing the action of censorship (or at least the programmatic texts that emanate from it), and of studying the treatises on poetics published in the sixteenth and seventeenth centuries. To prove the soundness of these intuitive remarks, born of literary sensibility and reading impressions, a quantitative study should be conducted.

The goal would be to compare different texts using machine learning techniques to determine which elements change over time and how these changes are consistent (if at all) with the intentions of the censors. It would be interesting to perform the same confrontations that were done intuitively in an automated way, proceeding in three steps: First, by comparing the original version of the Decameron with the three censored versions, to show the direction in which the censors' action changed Boccaccio's text; second, by comparing the original Italian novellas (by Boccaccio and Bandello in particular) with their versions translated into French, Spanish, and English, to see if the translations follow similar trends to those advocated by the censors; third, by comparing the short stories published before 1570 with those published after, to see how they evolve over time (and to compare this evolution with the censors' indications).

This research project is currently at the draft stage, and many questions remain unanswered, such as the best tools to speed up the OCR of Renaissance Italian, French, Spanish and English early prints¹⁸, and the multilingual nature of the corpus: how to compare texts written in different languages¹⁹, and even more so in ancient languages whose orthography is not yet fixed? Although we are still searching for the best way to answer these preliminary questions, we have tested the feasibility of this project and the possibility of modeling, through an automated survey, the changes that have taken place in novellas since the end of the sixteenth century. The test was carried out using some of the standard methods of quantitative analysis on a very small corpus. In fact, we compare only the sixteenth-century reference version of a short story by Boccaccio (i.e., the edition published by Giunti in Florence in 1527, generally called "ventisettana", which is the reference edition until the twentieth century, later mentioned as "Boccaccio (1527)") with the three censored versions already mentioned: the first by the Florentine academicians (called *Deputati*) published by Giunti in Florence in 1573 (later "Deputati (1573)"); the second by Lionardo Salviati (later "Salviati (1582)"), and the third written by Luigi Groto (later "Groto (1588)"). To begin this comparison, I have chosen one of the novellas most criticized by the censors for its irreligious and sensual plot, that is, the first novella of the third day, which tells the story of Masetto da Lamporecchio.

The story could be summarized as follows: Masetto da Lamporecchio hears that a convent of young nuns needs a gardener. He applied for the job, pretending to be mute. One day, in the garden, two nuns see Masetto sleeping, they desire him, they manage to sleep with him, thinking that he will not tell

¹⁶See Fumaroli 1985, p. 22-40.

¹⁷Patterson 1984, ch. 1.

¹⁸See <https://readcoop.eu/transkribus/>; <https://www.ocr4all.org/>.

¹⁹Some interesting research has been done to train an automatic translator in ancient English (Enrique Manjavacas Arevalo and Lauren Fonteyn, "2021, 23-36), or to normalize ancient Dutch (Marcel Bollmann 2012, 3-14). One of the problems is the lack of reliable digital libraries in ancient languages, Y. Hu, Z. LeBlanc, J. Diesner, T. Underwood, G. Layne-Worthey and J. S. Downie, 2022, 1-12. In France, there is one project of ancient digital library, concerning medieval French and English, see <http://palm.huma-num.fr/PALM/>.

anyone because he is mute. Soon all the nuns want to have sex with him, so Masetto is exhausted and can no longer work. He then decides to break his silence and explains to the Mother Abbess that he has regained his speech and wants to introduce less strenuous work rhythms. The nuns then organize a series of rollovers to have sex with Masetto without exhausting him, and they succeed in having Masetto appointed administrator of the monastery, with whom they live happily until Masetto retires, old, blessed with possessions and children (which he had from the nuns).

The three censored versions changed the original text in many ways that machine learning analysis could more accurately identify through unsupervised analysis.

3.1. Preliminary results: tokenization and word count

The first step necessary for any algorithmic analysis is the reduction of each of the considered novella versions to lists of *tokens*, or normalized, filtered words. Here we adopted the very simple (and unavoidably arbitrary) strategy to drop short words (conjunctions, prepositions, non-qualificative adverbs...), to ignore verbs and to map every remaining name and adjective to its unaltered, singular, orthograph-normalized form, retaining only gender (semantically relevant for our study). The lists of retained tokens from each version were then merged in a shared dictionary that included, in our case, $N_{\text{tokens}} = 321$ unique tokens.

The counting of token occurrences within each version is the simplest possible analysis that can then be performed after tokenization. Such a straightforward compilation already allows extracting some information. For instance, Table 1 lists 20 tokens appearing uniquely in the earliest Bocaccio (1527) version of the novella. These tokens are diverse but often refer to the Christian church domain, by direct reference to roles and actions within the monastery (badessa²⁰, *abbess*; monaca, *nun*; orazione, *prayer*...), religious garments (cocolla, *cowl*...) or directly Cristo, *Christ*. The absence of such tokens from later versions of the novellas suggests that the reduction of references to religion is going to be an important dimension of change. This initial analysis seems to confirm one of the results of our intuitive analysis of the short stories: the disappearance of the clerical character from the narrative and the laicization of the stories.

Many other differences, however, will be more subtle and involve multiple tokens shared by more than one version. Detecting them therefore requires analyses that go beyond counting individual words and look at the entire version-specific token count lists as a whole.

3.2. Latent Semantic Analysis shows a trend towards secularization and sentimental idealization

Our next step was thus to perform a standard *Latent Semantic Analysis* (LSA)²¹ to identify prominent dimensions of change in dictionary usage across the versions in a fully unsupervised manner, i.e. agnostically without a priori attention on pre-selected sensible tokens. In this analysis, the counts of occurrences of each token are seen as coordinates in a high dimensional space having as many dimensions as the number of considered tokens. Exactly as objects in ordinary three-dimensional space are characterized by providing three numbers (their width, height and depth), each text in the corpus is assigned a vector of N_{tokens} displacements in the N_{tokens} orthogonal directions of an abstract space of token counts. Visualizing such a space is a daunting task reminiscent of the adventures narrated in Edwin Abbot's *Flatland* Victorian novel, but machine learning techniques can be used to augment our limited geometric intuition. The LSA algorithm is thus used to identify a reduced number of dimensions (generally two or three for the sake of visualization) able to account in a more compact way for the variability in the original full-dimensional space. Once again, an analogy can be used to explain the concept of dimensional reduction. We live in a three-dimensional space, however we are used to specify the position of a city on Earth using just two numbers, the city's latitude and longitude.

²⁰We use a Courier font when referring to tokens, as their tokenized form (just an arbitrary symbol in a discrete list) is generally not present in the original texts, including instead multiple word forms normalized to this token.

²¹See S.T. Dumais, G.W. Furnas, T.K. Landauer and S. Deerwester, 1988, 281–285.

Table 1

Tokens appearing uniquely in the original Boccaccio (1527) version

Cristo	<i>Christ</i>
affare	<i>affair</i>
badessa	<i>abbess</i>
benda	<i>wimple</i> (headdress worn by nuns)
bianca	<i>white</i>
cappello	<i>hat</i>
cella	<i>cell</i> (of the monastery)
cocolla	<i>cowl</i> (long hooded garment worn by nuns)
cornia	<i>cuckhold</i>
famoso	<i>famous</i>
merito	<i>merit</i>
monaca	<i>nun</i>
monachino	<i>little monk</i>
monastero	<i>monastery</i>
nero	<i>black</i>
orazione	<i>prayer</i>
regione	<i>region</i>
santità	<i>holiness</i>
santo	<i>holy</i>
suora	<i>nun</i>

These two numbers do not correspond to horizontal and vertical displacement in an imaginary set of three Cartesian axes centered on the Earth's core, but are two angular non-Euclidean coordinates, specially adapted to account for the specific distribution of cities in 3D-space, bound to lay on the two-dimensional curved surface of the planet. Analogously, we can use LSA to construct *ad hoc* optimized coordinates able to capture the positional spread of the versions in the corpus in a more parsimonious but still faithful way. The method works better for large corpuses, when the large number of texts maps to denser point clouds in token count space providing a more systematic sampling of textual variability patterns. Nevertheless, it provides interpretable results already for our tiny corpus of four versions.

The results of LSA analysis on our corpus are shown by Figure 1. Each version of the novella corresponds to a dot in a scatter plot spanned by the three coordinates reconstructed by the LSA algorithm, known as *latent semantic factors*: latent because originally unknown; semantic because, by describing specific combinations of simultaneously occurring variations of word counts, these factors pretend capturing which topics are present in the overall corpus and to which extent these topics are addressed by each individual text in the corpus²². Each latent factor is precisely quantified by the degrees to which the counts of each specific token in the dictionary tend to increase or decrease when moving in the direction pointed to by the factor. Observing the positions in latent semantic space of the four Masetto novella versions, we immediately see that the latest Groto (1588) version stands apart as it is clearly displaced to the right, in the direction of the first latent factor, and to the back, in the direction of the second latent factor. These first two factors do not properly separate the Deputati (1573) and the Salviati (1582) versions from the earliest Boccaccio (1527) reference version (indeed their projections on the horizontal plane spanned by latent factors #1 and #2 do closely cluster). Yet, they more widely spread along the vertical direction of the third latent factor.

What is the meaning of these factors determined by the LSA algorithm as optimally describing the variability existing among the versions? Some of them are clearly interpretable. When moving in the direction of the second latent factor, counts of Religion-related tokens (as badessa, *abbess*; monastero,

²²The field of machine learning to which the LSA algorithm belongs is indeed sometimes called topic modelling.

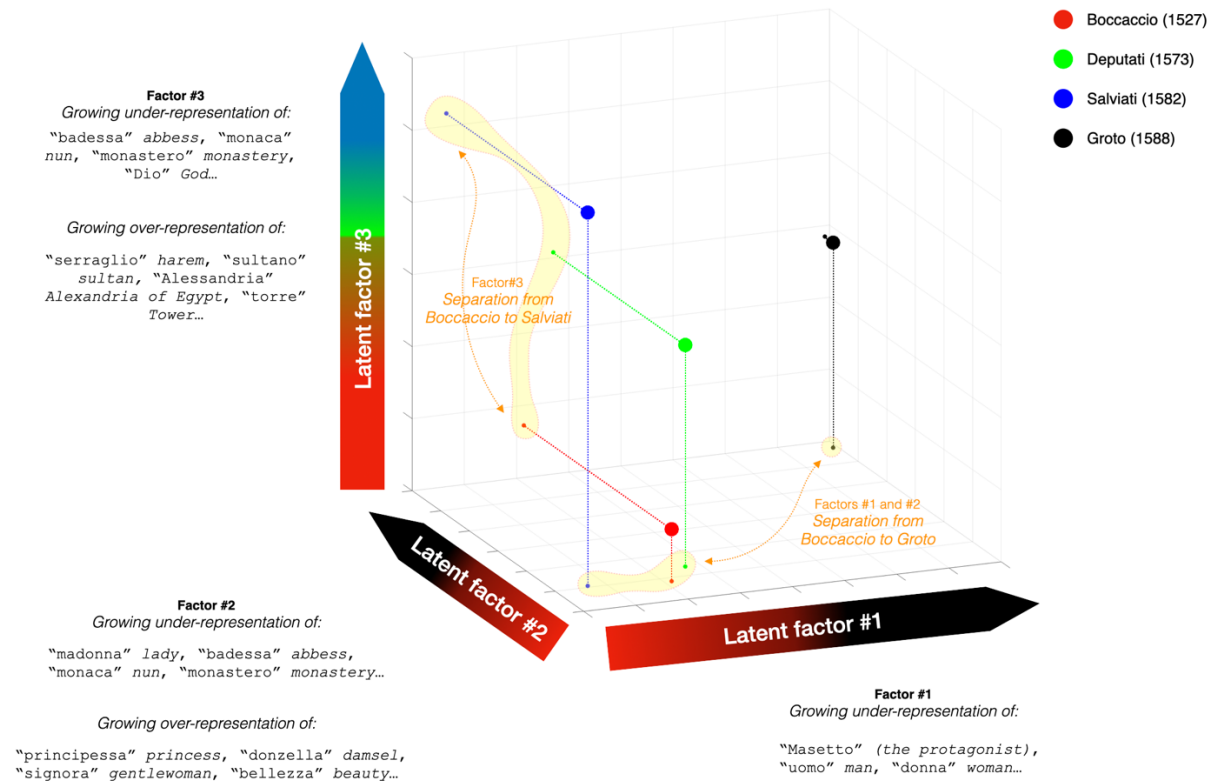


Figure 1: The latent space of diversity between the Masetto novella versions

monastery...) tend to decrease in parallel with the increase of the count of other tokens associated with courtliness and love (as *donzella*, *damsel*; *principessa*, *princess*; *bellezza*, *beauty*...). The second latent direction therefore seems to correspond both to a process of desacralization and to a convergence towards a more Romanesque style, as we move from the first versions to the last Groto edition (1588). The second latent direction thus confirms one of our intuitive hypotheses, namely that censorship led the novella to adopt the characteristics of the Greek novel, thus contributing to the successful spread of sentimental narratives, recounting the amorous adventures of idealized characters.

When moving in the direction of the third latent factor, the usual decrease of Religion-related tokens (including now even *Dio*, *God*) comes with an increase of tokens indicative of the shift towards exotic fictional contexts (as *serraglio*, *harem*; *sultano*, *sultan*; *torre*, *tower*...), which is indeed progressively occurring in the "Deputati (1573)" and "Salviati (1582)" versions. The third latent factor reflects thus again a process of laicization and shows a tendency to move the narrative to exotic countries. This trend supports the hypothesis that the novellas are getting closer to the Greek novel (which is often set in oriental countries) and reveals two important features of censorship. The displacement of the narrative to a distant (and pagan) land removes any risk of blasphemy, since the characters are no longer Christians. This shift results in a greater diversity of characters, many of whom are then clearly racialized.

Surprisingly, the first latent factor, which is the one capturing the largest amount of statistical variation, happens to be the least interpretable among the three considered. This factor corresponds to the decrease in the occurrence of rather general words, the most prominent of which being associated with a tendency in "Groto (1588)" to more frequently avoid the direct use of the proper name of the protagonist, Masetto. The inclusion of more texts in the corpus would be required to confirm the existence of such a tendency and attribute it to general stylistic or linguistic changes.

The LSA algorithm is scalable and flexible and can be generalized to much wider corpuses and dictionaries, including not only individual tokens, as in Figure 1, but also "*k*-grams", or lemmatized text chunks of *k* successive tokens. Increasing the order *k* of the considered text chunks allows in principle

an increasingly accurate topic modelling and is also the pre-requisite for the perspective training of unsupervised language models able to develop self-organized attention to semantically sensible terms²³. Here, however, the too limited corpus size prevents extending LSA modelling beyond k -grams composed of more than two tokens (bi-grams). As reported in Table 2, a second order LSA analysis confirms some of the trends already revealed by the first order analysis of Figure 1, by adding some additional nuances.

Table 2

Example bi-grams growingly over-or under-represented in the latent directions separating Boccaccio (1527) from Groto (1588) versions

Growing under-representation		Growing over-representation	
buono uomo	<i>good man</i>	femmina femminile	<i>womanly woman</i>
povero uomo	<i>poor man</i>	femminile	<i>womanly appetite</i>
		appetito	
uomo femmina	<i>man woman</i>	turbamento	<i>agitation against nature</i>
		contra-natura	
Masetto	<i>Masetto Lamporecchio</i> (the	concupiscente	<i>lustful appetite</i>
Lamporecchio	full name of the	appetito	
	protagonist)		

A tendency to reduce plain references to male characters (buono uomo, povero uomo, *good man*, *poor man*...) or direct use of proper nouns (Masetto Lamporecchio) is again retrieved, interestingly coupled to an enhanced reference to feminine qualities and sensual aspects (femmina femminile, “*womanly woman*”; concupiscente appetito, *lustful appetite*...). This trend would need to be studied on a large corpus, but in the present case it hints at an evolution in censored texts towards greater attention to the sensual and sentimental elements of the story. In other words, we can assume that while the censors remove all explicit references to sex, they add erotic and sentimental allusions.

A slightly modified LSA algorithm can also operate on sets of split sentences. This is what we show in Figure 2, where every dot does not correspond now to a whole novella version as in Figure 1, but to individual sentences within the novellas (dots are colored according to the version from which they are extracted). Rather than modelling a data cloud of four points we can thus parse a larger-size set of 244 sentences. This allows a finer modelling, now in terms of 30 latent factors rather than just three (although we can only visualize in the plot the projections of the sentences in the space spanned by the first three latent factors). Every sentence maps to a point in an abstract space and we can thus measure the pairwise distances between sentences using the generalization of the standard Euclidean distance to this space. Some sentences will be more central (that is closer, on average, to any other sentence, denoting a form of topical sharedness across the corpus), while other will be more eccentric (that is, more distant from the rest of other sentences, reflecting a form of rarer topical exceptionality)²⁴. The size of the dots in Figure 2 corresponds to their increasing eccentricity.

We highlight specifically the four most eccentric sentences, which eventually are highly characteristic passages of the novella version to which they belong. Two excerpts from the reference version “Boccaccio (1527)” stand out. The first explicitly describes the sexual desire that persists in nuns, who “does not become of stone” even if “a white wimple is put on her head and a black cowl over her body”²⁵. The second excerpt describes the immoral trick played by the nuns, who decide to lie and cheat in order to keep Masetto in the convent: they tell the locals that, thanks to their prayers, Masetto, “who had been

²³See e.g. the word2vec embedding framework (T. Mikolov, et al. 2013) and its many generalizations, lying also behind sophisticated language models as the popular ChatGPT.

²⁴We cannot give full methodological detail here. In short, we derive from the LSA embedding a matrix of pairwise distances and construct from it a suitably weighed graph on which we can perform closeness analyses (see Newman, M.E.J. Networks: An Introduction. Oxford, UK: Oxford University Press (2010), chapters 6 and 7).

²⁵“There are some men and women who are foolish enough to believe that when a young woman has a white wimple put on her head and a black cowl on her body, she is no longer a woman and no longer has female desires, as if becoming a nun had turned her into a stone” (my translation).

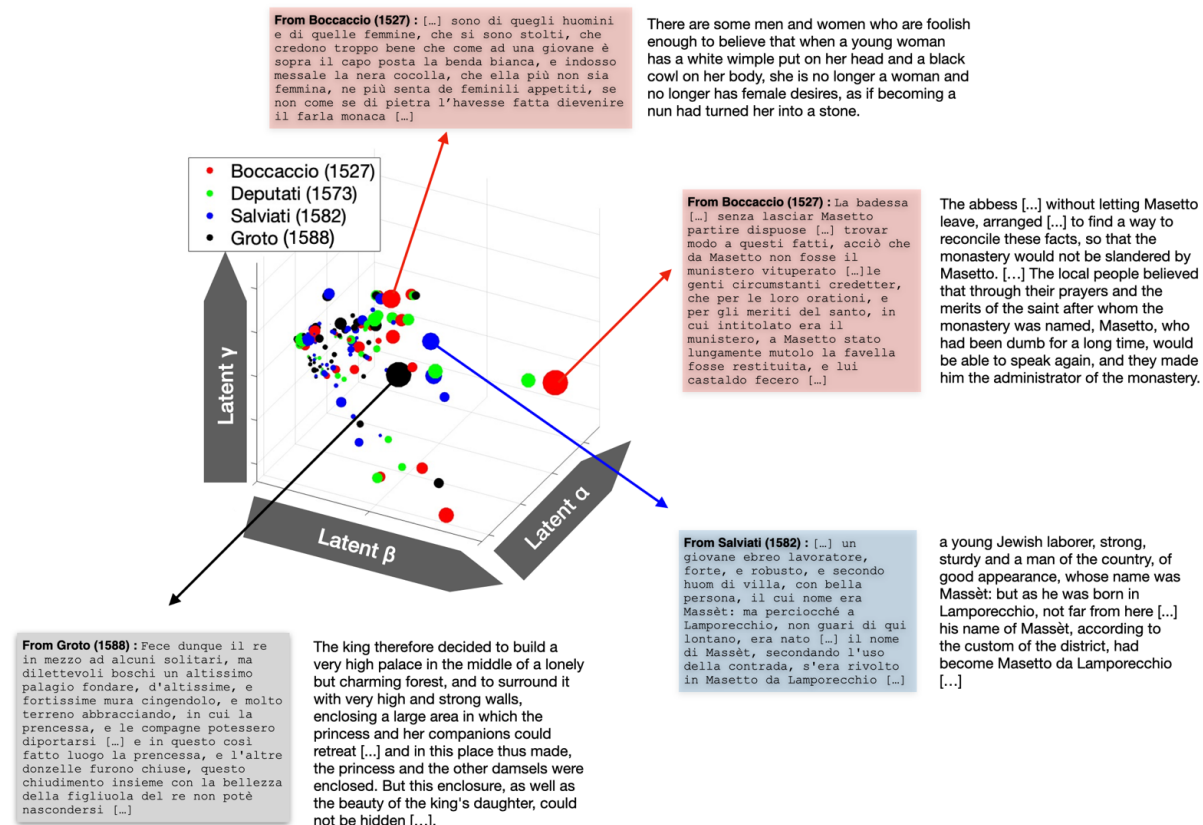


Figure 2: Eccentricity of sentences in the Masetto novella versions

mute for a long time, would be able to speak again”, and persuade them “to make him the administrator of the convent”. Thanks to this trick, the nuns are free to satisfy their sexual desires and to arrange shifts and a schedule to have sex with Masetto as long as they can²⁶. These two sentences were censored in later editions, and it is easy to see why. The second one is particularly interesting because it goes against poetic justice: instead of being punished for their several sins (adultery, blasphemy, lying...), Masetto and the nun can easily deceive the locals and live happily ever after. On this account, it confirms one of our intuitive hypotheses that the censored version tries to alter the plot to restore a form of poetic justice (or at least to erase any kind of explicit poetic *in-justice*).

Other excerpts pop out for their introduction of fictional elements (a princess secluded with her damsels in a high castle from “Groto (1588)”²⁷) or for the exotic racialization of certain characters: in “Salviati (1582)”, Masetto become a Jew²⁸. Two reasons could explain the racialization of Masetto in Salviati’s version: since Salviati moves the setting of the story to Egypt, the characters needs to change origins and religion. Moreover, since Jews in the early modern period are generally accused of duplicity, Masetto’s scheme (pretending to be mute to be employed by the nuns and eventually satisfying their sexual desires) can be explained and condemned through a racial prejudice. Salviati’s version of the

²⁶“The abbess [...] without letting Masetto leave, arranged [...] to find a way to reconcile these facts, so that the monastery would not be slandered by Masetto. [...] The local people believed that through their prayers and the merits of the saint after whom the monastery was named, Masetto, who had been dumb for a long time, would be able to speak again, and they made him the administrator of the monastery” (my translation).

²⁷“The king therefore decided to build a very high palace in the middle of a lonely but charming forest, and to surround it with very high and strong walls, enclosing a large area in which the princess and her companions could retreat [...] and in this place thus made, the princess and the other damsels were enclosed. But this enclosure, as well as the beauty of the king’s daughter, could not be hidden” (my translation).

²⁸“[...] a young Jewish laborer, strong, sturdy and a man of the country, of good appearance, whose name was Massèt: but as he was born in Lamporecchio, not far from here [...] his name of Massèt, according to the custom of the district, had become Masetto da Lamporecchio”(my translation).

novella thus reaffirms contemporary racial prejudices, thus confirming one of our hypotheses: that the censored version tends to condemn characters as racially prejudiced.

3.3. Modeling change through text sequence analysis, inspired by genetics

These examples show that modelling in terms of latent semantic factors or other word embeddings are promising tools for the unsupervised and completely data-driven detection of trends ruling textual changes across multiple editions. A limitation of these methods, however, is that they consider documents as “bag of words”, using a technical term common in the machine learning literature. Eventually, novellas are sequentially organized narratives in which the order and precise position of words and sentences, beyond their presence, may also affect meaning and thus become a potential target for censorship interventions. Additional analysis tools are thus needed to detect differences as well at the level of precisely ordered sequences of tokens. From an algorithmic point of view, the comparison of the sequences of tokens associated to two different versions of a same novella is very close to the problem of gene sequence alignment in bioinformatics. Homologous DNA sequences coding for important proteins are very similar across species (e.g. human vs whale or worm hemoglobin) and all share highly preserved chunks coding for conformational elements necessary to the protein’s function (e.g. fixing and releasing oxygen molecules). However, these chunks do not appear precisely at the same position within the sequence, as different sequences may have different lengths due to the absence of certain elements (worms hemoglobin is more primitive than mammals’, reflecting phylogeny of species) or the insertion of additional elements, sometimes as replacement of others (whale hemoglobin is modified relatively to human, because it can fix a larger number of oxygen molecules, an adaptive evolution desirable for a mammal colonizing the ocean). This is not dissimilar from textual changes across editions where entire sections can be added to please a different public and individual “point mutations” (word removals or replacements) arise as an effect of censorship interventions or as a protection against them. The genomes of living species evolve under the action of natural selection to allow survival and spread in changing natural environments. Analogously, novellas would be modified by editorial interventions in order to remain compatible with changing cultural and political contexts.

This analogy between the evolution of biological and textual sequences have already been exploited in linguistic computing for the automated detection of plagiarism²⁹ and intertextuality³⁰ or for stemmatological applications to uncover the relationships between different manuscript versions of a text, as Chaucer’s *Canterbury Tales*³¹. Here we extend it to compare the four versions of the Masetto novella, where changes may be more extensive and have been driven by directed editorial intention rather than by errors in copying. Specifically, we compare the sequences of tokens of each of the four versions, so that differences at the level of orthographic and grammatical variants or short words are ignored, focusing instead on substitutions, insertions, and deletions of semantic elements. The results of the sequence comparisons are shown in Figure 3.

In these sequence alignment graphs the first row always corresponds to the reference version “Boccaccio (1527)” and the second row to one of the three later editions. Different yellow positions from left to right correspond to different tokens along the textual sequence of the novella. The blue color corresponds to mismatched tokens, as some tokens present in one version (in yellow in the corresponding row) do not have a matching counterpart in the other row (hence the blue gap). Such gaps are automatically inserted by the sequence alignment algorithm³², in order to maximize the match between the tokens of the reference and the modified versions. Thus, a blue gap in the first row corresponds to the insertion of tokens in the later version that were not present in the original version, while a blue gap in the second row corresponds to the deletion of tokens from the reference version that are no longer present in the later version. Sequence alignment graphs thus provide a bird’s eye

²⁹See e.g. C.K. Ryu, et al. 2008, 119–124.

³⁰See e.g. R. Horton, R., M. Olsen, and G. Roe, 2010 (online: <https://www.digitalstudies.org/article/id/7224/>), studying the reuse of old texts in Diderot and D’Alembert Encyclopédie.

³¹See C.J. Howe, et al. 2001, 147–152. For a review and, for a specific application, see M. Spencer, et al. 2003, 407–422.

³²A simple adaptation of a classic algorithm in bioinformatics, S.B. Needleman and C.D. Wunsch, C.D. 1970, 443–53.

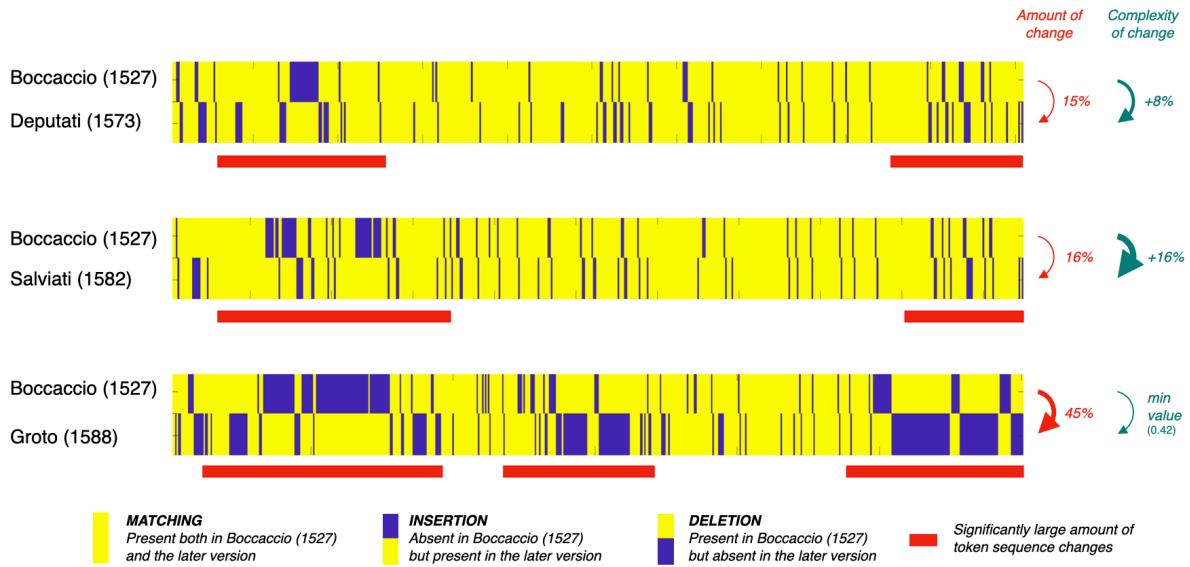


Figure 3: Changes in the token sequences in the Masetto novella versions

view of the extent of differences between two versions and where in their textual sequence changes tend to occur.

A first quantity that can be extracted from these graphs is the total amount of change. The 15% and 16% of the token sequence positions appear modified, respectively, in the “Deputati (1573)” and the “Salviati (1582)” versions relatively to the reference “Boccaccio (1527)”. As visually evident from the larger amount of blue color, the amount of change in “Groto (1588)” is way larger, reaching the 45% of token positions. The type of modification interventions is also different. The number of individual edits is similar across the three versions (we can count 41 deletions and 33 insertions on average). However, the lengths of these individual editing interventions are different. In the “Deputati (1573)” and the “Salviati (1582)” most of the editing interventions are very short and surgical, involving deletion of at most two tokens in the 85% and the 93% of cases, respectively, and insertions of at most two tokens in the 89% of cases. In contrast, in “Groto (1588)” the fractions of short deletions and insertion drop to 71% and 65% and more than one quarter of the edits correspond to the insertion or removal of entire sentences (i.e., larger than four tokens, remembering that tokenization ignores linking words and verbs). The top 5% of longest deletions and insertions in “Groto (1588)” are longer than 20 tokens and go up to 37 (entire paragraphs), while none of the edits in “Deputati (1573)” and “Salviati (1582)” go beyond 10 tokens (and 95% of changes are shorter than four tokens). In all cases, most changes tend to concentrate in the beginning and the ending of the novella, which set the overall frame for the story and enforce a more or less comic or moral conclusion³³. This result seems to show that the rewriting action is not only focused on singular words (abbess, nun, etc.), but also on the setting of the novella, which is radically changed in the censored versions, no longer taking place in a convent, but in a house for young girls (Deputati 1573), in a harem (Salviati 1582), or in the magnificent castle of a princess isolated in the woods (Groto 1588). Our intuitive hypothesis that the censors adopted a “myopic” reading of the text therefore needs to be revisited in order to better understand reading and rewriting strategies.

Besides and beyond the evaluation of the amount of changes, it is also possible to evaluate the computational complexity of generating the observed changes. Such an abstract quantity was defined by the Russian mathematician Andrei Kolmogorov as the minimum number of instructions that must be given to a mechanical agent in order for it to produce a sequence of symbols³⁴. Here, the process of

³³The enhanced concentration of edits is indicated in Figure 3 by red segments under specific sequence intervals. These intervals are estimated statistically using a permutation testing procedure (95% of statistical significance).

³⁴For a general introduction to the topic of symbolic sequence complexity a first read could be G. Chaitin, *Exploring Randomness*, Berlin, Springer Verlag, 2001. Kolmogorov first introduced the notion of sequence complexity to study the complexity of

producing an edited version can be modeled as the computational procedure of scanning the original “Boccaccio (1527)” version token by token and deciding whether to keep the token, discard it, and/or insert a new token to produce another version. If we assume, as a first approximation, that each of these operations has the same “cost” to the editing agent, then the effort required to generate an edited version can be evaluated as the computational complexity of writing each of the sequence alignment graphs of Figure 3. Applying this analysis to the case of the three censored versions, i.e. “Deputati (1573)”, “Salviati (1582)” and “Groto (1588)”, we find that the complexity of generating a version does not necessarily parallel the total amount of changes that this version contains. The “Groto (1588)” version contains the largest amount of changes, but the effort required to produce these changes, according to our symbolic manipulation model, is 8% less than that required to produce the “Deputati (1573)” version, and 16% less than that required to produce the “Salviati (1582)” version. Our analysis procedure is abstract and its interpretation is not immediately obvious. Nevertheless, we can speculate that the lower complexity of editing in “Groto (1588)” manifests its more drastic rewriting, while “Deputati (1573)” and “Salviati (1582)” attempted to remain more faithful to the original version, while still achieving semantic reorientations with a minimum of intervention. It is also intriguing that the most “laborious” version, Salviati’s edition of 1582, was also the most successful of the three, being reedited several times and becoming the new standard version of the Decameron for at least two centuries.

It would be useless and arbitrary to extrapolate more conclusions given the very limited size and *ad hoc* nature of our corpus. Yet, these first preliminary computational explorations provide a promising proof-of-concept that machine learning tools can be used to achieve a fine modelling of the effects of an increasing and thematically evolving censorship pressure. It is interesting to note that the unsupervised results, obtained by standard latent semantic analysis (applied to tokenized single words, token pairs, or sentences) or by sequence analysis inspired by genetic methods, partially confirm our intuitive hypothesis, and namely that of the progressive secularization and racialization of the novella, that of the moralization of its plot and its revision according to the logic of poetic justice, and that of the idealization of the plot and its transfer to a remote or imaginary place in the style of the sentimental novel. The unsupervised analysis does not confirm the hypothesis of a censorial “myopic reading”, but rather shows that the reading strategy promoted by censorship needs to be further investigated. The theoretical idea that censorship promoted the rise of the verisimilar narrative and thus contributed to the spread of Aristotelian poetics seems difficult to prove with quantitative analysis and needs to be further explored and observed in the narrative forms.

The use of much larger corpuses will open the way to the use of more advanced methods, even if a tradeoff between enhanced precision of modelling and maintenance of interpretability will always have to be considered. These exploratory analyses show the value of studying the changes in literary practices at the end of the Renaissance with quantitative methods, in the hope of defining with a reliable model what is for the moment no more than a reading impression, a metaphor (“the fall” of the Renaissance), or a theoretical hypothesis (“the crisis of exemplarity”), and thus being able to create a more reliable model of change in literary history.

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepL write in order to check the Grammar and spelling. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

Russian versification rhythms in A. Kolmogorov and N. Rychkova, “Russian Poetry Rhythm Analysis and Probability Theory” (translated from the Russian by Rosina, R.), *Theory of Probability and Its Applications*, n. 44(2), 1962 (posthumous publication in 2000), p. 375–385. Only later he developed the rigorous general-purpose formalism today known as Kolmogorov-Chaitin complexity theory.

References

- [1] S. Bargagli, *I trattenimenti*, ed. L. Riccò, Salerno, Roma, 1989.
- [2] G. Boccaccio, *Decameron*, ed. « deputati fiorentini », Firenze, Giunti, 1573.
- [3] G. Boccaccio, *Decameron*, ed. Lionardo Salviati, Firenze, Giunti, 1582.
- [4] G. Boccaccio, *Decameron*, ed. Luigi Groto, Venezia, Farri, 1588.
- [5] M. Bollmann, "Semi-Automatic Normalization of Historical Texts using Distance Measures and the Norma tool", in *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, Lisbon, Portugal (2012): 3-14.
- [6] J. Martínez de Bujanda (ed), *Index de Rome, 1590, 1593, 1596*, University of Sherbrooke, Droz, 1994.
- [7] S. Carrai and S. Madricardo, "I Decameron censurato: preliminari alla "rassettatura" del 1573", *Rivista di letteratura italiana*, 7 (1989): 225-247.
- [8] G. Chiecchi (ed), *Dolcemente dissimulando : cartelle laurenziane e decameron censurato 1573*, Antenore, Padova, 1992.
- [9] C. S. Clegg, *Press Censorship in Elizabethan England*, Cambridge University Press, Cambridge, 1997.
- [10] N. Crémona, *Poétique des histoires tragiques (1559-1644), «Pleines de chair et de sang »*, Garnier, Paris, 2019.
- [11] S.T. Dumais, G.W. Furnas, T.K. Landauer and S. Deerwester, "Using latent semantic analysis to improve information retrieval", *Proceedings of CHI'88 Conference on Human Factors in Computing Systems*, ed. J.J O'Hare, New York, Association for Computing Machinery (1988): 281-285.
- [12] G. Fragnito, *Rinascimento perduto, la letteratura italiana sotto gli occhi dei censori (secoli xv-xviii)*, Il Mulino, Bologna, 2019.
- [13] V. Frajese, *Nascita dell'Indice : la censura ecclesiastica dal Rinascimento alla Controriforma*, Morcelliana, Milano, 2008 (2006).
- [14] M. Fumaroli, "Jacques Amyot and the Clerical Polemic Against the Chivalric Novel", *Renaissance Quarterly*, 38, 1 (1985): 22-40.
- [15] A. Hadfield (ed.), *Literature and Censorship in Renaissance England*, Palgrave, London, 2001.
- [16] R. Horton, R., M. Olsen, and G. Roe, "Something Borrowed: Sequence Alignment and the Detection of Similar Passages in Large Text Collections". *Digital Studies – Le Champ numérique*, 2.1 (2010), online: <https://www.digitalstudies.org/article/id/7224/>
- [17] C.J. Howe, et al. "Manuscript Evolution", *Trends in Genetics*, 17/3, (2001): 147-152.
- [18] Y. Hu, Z. LeBlanc, J. Diesner, T. Underwood, G. Layne-Worthey and J. S. Downie, "Complexities Associated with User-generated Book Reviews in Digital Libraries: Temporal, Cultural, and Political Case Studies 2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Cologne, Germany (2022): 1-12.
- [19] S. Landi, *Stampa, censura e opinione pubblica in età moderna*, Il Mulino, Bologna. 2011.
- [20] E. Manjavacas Arevalo and L. Fonteyn, "MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950)" in *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, NIT Silchar, India, (2021):23-36.
- [21] G. Minois, *Censure et culture sous l'ancien régime*, Fayard, Paris, 1995.
- [22] B. Needleman and C.D. Wunsch, C.D., "A general method applicable to the search for similarities in the amino acid sequence of two proteins", *Journal of Molecular Biology*, 48/3 (1970): 443-53.
- [23] A. Patterson, *Censorship and Interpretation, the Conditions of Writing and Reading in Early Modern England*, Madison, University of Wisconsin Press, 1984.
- [24] V. Pinto Crespo, *Inquisición y control ideológico en la España del siglo xvi*, Tauris, Madrid, 1983.
- [25] B. Poissonot, *Nouvelles histoires tragiques*, ed. Jean-Claude Arnould and Richard A. Carr, Genève, Droz, 1996.
- [26] W. K. Pietrzak, *Le Tragique dans les nouvelles exemplaires en France au xvie siècle*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, 2006.
- [27] L. Plazenet, *L'Ébahissement et la délectation. Réception comparée et poétique du roman grec en France et en Angleterre aux xvi^e et xvii^e siècles*, Champion, Paris, 1997.

- [28] Ryu, et al. (2008) "Detecting and Tracing Plagiarized Documents by Reconstruction Plagiarism-Evolution Tree". In IEEE International Conference on Computer and Information Technology, (2008): 119–124.
- [29] T. Rymer, *The Tragedies of the Last Age*, Richard Tonson, London, 1678.
- [30] R. Smythe, *Strange Lamentable and Tragicall Hystories*, Hugh Jackson, London, 1577.
- [31] M. Spencer, et al. (2003) "How Reliable is a Stemma? An analysis of Chaucer's Miller's Tale", *Literary and Linguistic Computing*, 18/4, (2003): 407-422.