



HAL
open science

The Authorship of the Works of Chrétien de Troyes: a Stylometric Examination

Jean-Baptiste Camps, Benedetta Salvati, Gonzalo Freijedo, Donghan Bian, Gaëtan Drouet, Eglantine Gaglione, Émilie Guidi, Carolina Macedo, Yaelle Zribi, Florian Cafiero

► **To cite this version:**

Jean-Baptiste Camps, Benedetta Salvati, Gonzalo Freijedo, Donghan Bian, Gaëtan Drouet, et al.. The Authorship of the Works of Chrétien de Troyes: a Stylometric Examination. DH Benelux 2024, 2024. hal-04606835

HAL Id: hal-04606835

<https://enc.hal.science/hal-04606835>

Submitted on 10 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

The Authorship of the Works of Chrétien de Troyes: a Stylometric Examination

Jean-Baptiste Camps¹, Benedetta Salvati¹, Gonzalo Freijedo Aduna¹, Donghan Bian¹, Gaëtan Drouet¹, Eglantine Gaglione¹, Émilie Guidi¹, Carolina Macedo¹, Yaelle Zribi¹, and Florian Cafiero¹

¹École nationale des chartes, Université PSL

1. Introduction

Chrétien de Troyes, a pivotal and inspirational figure in medieval French literature (Kibler, 2007), played a crucial role in shaping the Arthurian romance genre. Despite his prominent role and heritage, little is known of Chrétien’s life. Active in the late 12th century, he was a poet for the court of Marie de Champagne (Doudet, 2009). The lack of reliable sources about Chrétien has allowed numerous controversies to arise regarding the delimitation of his works. A sentence at the end of *Lancelot ou le Chevalier de la charrette* states that a Godefroi de Leigni would have finished the text, but the extent of the work of this unknown author remains somewhat unclear, despite previous interesting quantitative analyses (Reilly and Dillon, 2013). The romance *Guillaume d’Angleterre*, signed under the name “Crestiens” has sometimes been attributed to Chrétien de Troyes, fostering a decades-long debate (Slautina, 2012, Williams, 1987, Zufferey, 2008), and is considered by some as certainly not a work by Chrétien de Troyes (Lacy and Grimbert, 2005, Walter, 1997). Some specialists have suggested that the style used in Chrétien de Troyes’ supposed adaptations of Ovid (*Philomena*) is too different from the style for which he is known (Krueger, 2005).

Finally, the different *Continuations* to Chrétien’s *Perceval* are also a topic in themselves, having been variously attributed to Chrétien himself, Wauchier de Denain (Second Continuation), or Gerbert de Montreuil (Fourth Continuation), among other hypotheses (Bruckner, 2006). Though we include Gerbert de Montreuil’s alleged Continuation in our corpus, we set aside the question of the authorship of all continuations as the topic of a future investigation.

In this paper, we try to address these debates using stylometry, an approach that has proven useful to attribute literary texts, even in the case of collaborations or sequential authorship (Cafiero and Camps, 2021, Plecháč, 2021), and in the complex context of medieval texts (Camps et al., 2021b, Kestemont et al., 2015, Van Dalen-Oskam and Van Zundert, 2007). We build a corpus by breaking the silos between different data sources, and then subject it to both unsupervised and supervised analysis, the latter using a rolling approach suited to the suspicions of collaborative authorship present in Chrétien’s later works.

2. Material and methods

2.1. Corpus and Data

The corpus was built around the works of Chrétien de Troyes (fl. c. 1170-1190) himself, including all his certainly attributed romances, as well as the *Lancelot*, started by Chrétien and supposedly finished by another author, Godefroi de Leigni. The disputed works of *Philomena* and *Guillaume d'Angleterre* were included in order to be analysed.

Two imposters were then added: Gerbert de Montreuil (fl. first half of the 13th c.) was chosen, as the author of a *Continuation* to Chrétien's works, and therefore producer of texts very close in terms of genre, form and content. Finally, Jean Renart (fl. end of the 12th-beg. of the 13th c.) was included as an outsider.

To that end, we have compiled data from various sources to design a corpus: the works of Chrétien reused the data of a digital edition of the copy made by the scribe *Guiot de Provins* (Kunstmann, 2009); other texts were selected from the *Base de français médiéval* (Guillot et al., 2018), that provides the text of digitised editions. Finally, we completed the corpus by performing OCR and post-correction on some texts for which no digital editions were available, using eScriptorium (Kiessling et al., 2019) and the model *CATMUS Print* (Gabay et al., 2024). The detailed composition of the corpus is presented in Appendix A.

Lemmatization was performed using Pie (Manjavacas et al., 2019) and a model trained specifically for Old French (Camps et al., 2020, 2021a)¹.

Function words were then extracted on the lemmatised text, using a custom list, built from the lemmatised Gold data of the 0F3C Corpus (Camps et al., 2021a).

2.2. Stylometric Analysis

To build a training corpus, we carefully select works that appear mostly free of doubts regarding attribution. In particular, for Chrétien de Troyes, his two earlier kept romances were selected, under the assumption that they are the ones with the least risk of collaborative authorship. Indeed, there is no mention of such collaboration before *Lancelot*, nor are there known *Continuations*.

Yet, since medieval texts are rarely free of doubts about authorship, and before proceeding to a supervised analysis, we still need to check the consistency of the traditional attribution of the works of our corpus – that is, their homogeneity in terms of style and attribution.

For this, we selected a potential training corpus of two works by each author (Chrétien de Troyes, Gerbert de Montreuil and Jean Renart) and submitted it to unsupervised analysis. If this step confirms their homogeneous nature, we can then proceed to using them as training material in the next step, using SVM to perform rolling analyses on the texts.

In both cases, we sample the texts in 3000 words chunks, and then extract lemmata frequencies of function words.

2.2.1. Unsupervised analysis

Data is normalised, and distance computed using a combination of Burrow's delta and L2-norm, and then hierarchical clustering is applied, using Ward's method (Cafiero

¹ This model actually includes works from Chrétien de Troyes in the training data, which can in the same time improve its efficiency on our corpus, and create biases between the lemmatisation success of the works of this author and the rest.

and Camps, 2019, Evert et al., 2017).

2.2.2. Supervised analysis

We then train support vector machines, with a linear kernel, using group-k-fold as a cross-validation strategy (i.e., each work is iteratively removed from the training data and its samples used as a test set, before a global score is computed). The computations were performed using the software Superstyl (Camps, 2021). Results of training are presented in Table 1. The most important contributing features for each classifier are provided in Appendix B.

Table 1: Classification Metrics

Class	Precision	Recall	F1-Score	Support
ChrTr	0.81	0.81	0.81	26
GerbM	1.00	0.91	0.95	23
JnRen	0.81	0.87	0.84	30
<i>accuracy</i>			0.86	79
<i>macro avg</i>	0.87	0.86	0.87	79
<i>weighted avg</i>	0.87	0.86	0.86	79

3. Results

Hierarchical clustering results (Figure 1) can be cut into three main clusters. One contains all samples from works attributed to Gerbert de Montreuil, the second all samples from works of Jean Renart, plus two samples drawn from *Cliges* and *Erec et Enide*. The third one contains all the remaining samples from the works attributed to Chrétien de Troyes, as well as *Philomena* and *Guillaume d’Angleterre*. Most samples from this last work are clustered together in a sub-branch, at the frontier between the works of Chrétien de Troyes and those of Jean Renart.

The results of rolling analyses are shown in Figure 2. In both *Lancelot* and *Perceval* there is a significant drop towards the end in Chrétien de Troyes’ curve (representing the value of the SVM classifier for Chrétien’s style), and increase in Gerbert’s (who is even shown as a better candidate for some limited portions, but without a value > 0).

In both *Philomena* and *Guillaume d’Angleterre*, Chrétien globally scores the highest, but with relatively low values concerning the last of these texts, apart from a section representing roughly the third quarter of this work, in which values similar to those observed for *Yvain* are found.

4. Discussion

Our results can be arranged in three groups: first, some of them seem to confirm current consensus on the authorship of these works; secondly, another group contains potentially new findings, even on works until now not debated as far as we know; and thirdly, there are some partially inconclusive results, that call for further investigation.

Firstly, the unsupervised analysis seem to globally confirm the coherence of three groups of works, usually attributed to, respectively Chrétien de Troyes, Gerbert de Montreuil (*Violette* and Fourth Continuation) and Jean Renart.

In what concerns suspicions of collaborative authorship in Chrétien de Troyes’ later works, our rolling analysis seems to confirm a significant drop in Chrétien’s stylistic

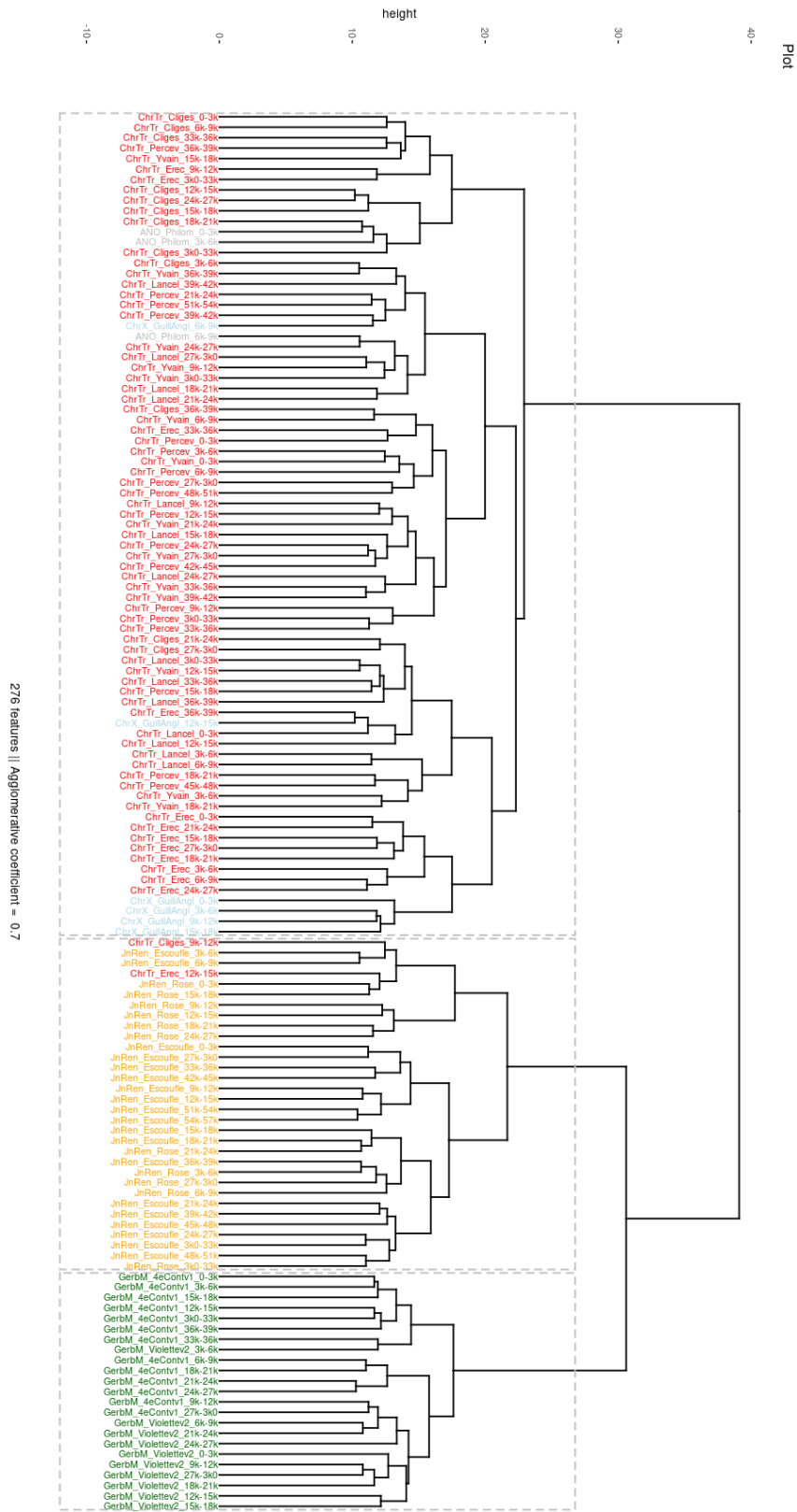


Figure 1: Hierarchical clustering of the corpus (Ward's method, Burrows' delta and L2 norm)

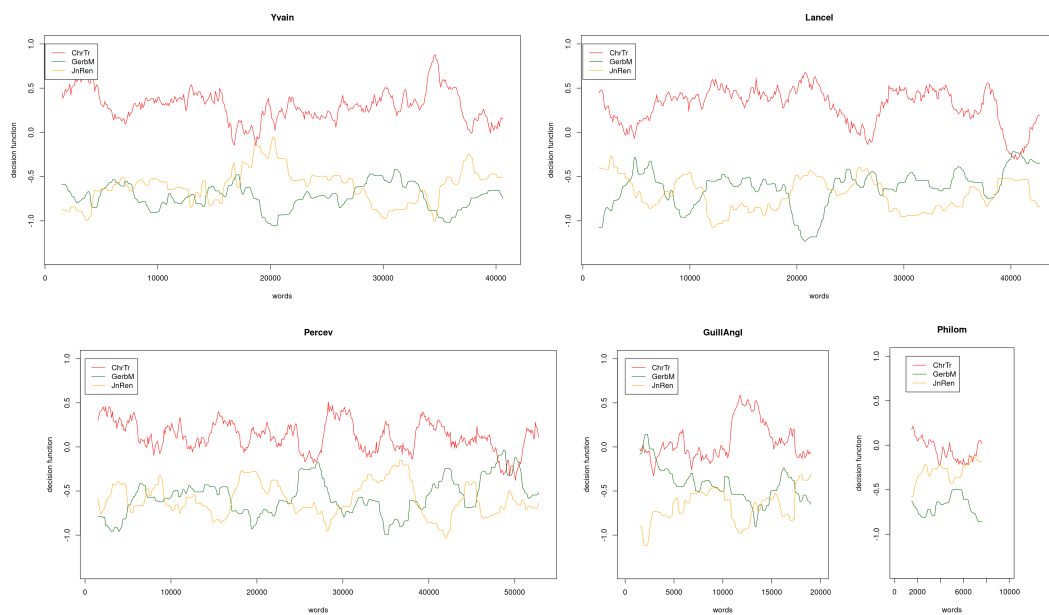


Figure 2: Rolling SVM analyses of all texts unseen in training

signature at the end of *Lancelot*, roughly from word 38 200 in our analysis, which matches the verse 6156 in our edition (“*Mes de Lancelot n’i voi mie*”). This is quite coherent with both the stylistic expertise of James-Raoul (2007) and the quantitative analysis of Reilly and Dillon (2013), who posit a transition from Chrétien de Troyes to Godefroi de Leigni around verse 6150.

But there are results on collaborative authorship that are also novel. In particular, there is at least one drop similar to the one observable in *Lancelot* (roughly v. 6150-end), but this time at the end of *Perceval*. It spans from word 47 500 to 51 000, that is, roughly, from verse 7841 (“*Lors s’est la reine esmeüe*”) to 8416 (“*qu’ausinc orroie ge conter*”), before Chrétien’s signal takes over again. This is quite an interesting finding inasmuch as *Perceval* is known to be an unfinished work, probably interrupted by Chrétien’s death, that has spurred the contribution of several continuators. Is it possible too that, like for *Lancelot*, Chrétien obtained the contribution of an anonymous hand towards the end of the text. Or, perhaps, are we faced with the presence of a posthumous interpolation?² In both cases, the curves goes up slightly again at the very end, perhaps indicating that the last few verses are in Chrétien’s own writing.

Regarding the case of *Philomena* and *Guillaume d’Angleterre*, our current results call for further investigation. In the case of *Philomena*, both the unsupervised and the rolling analyses seem to confirm that *Chrétien* is globally the best candidate in our set, but without a strong certainty.

The case of *Guillaume d’Angleterre* apparently calls for nuance: both analyses show that some portions of this work are quite close to Chrétien de Troyes’ style, while other are more autonomous. In particular, in the rolling analysis, the portion from word 10 500 to 14 500 (roughly, from v. 1692, ‘*Et Loviax s’est tos resjois*’, to 2332, ‘*S’il vos*

² Interestingly, the end of this portion matches a response from Grinomalanz to Gauvain, in which he contests the veracity of the Gauvain’s narrative in these words: “Or m’est il solaz et deliz / de tes mançonges escouter, / qu’ausinc orroie ge conter / .i. fableor con ge faz toi. / Tu es juglerres, bien le voi.” (*It is to me agreeable and pleasurable to listen to your lies, in the same way as I would hear a composer of fiction telling a tale. I can tell you are a minstrel*).

venoit a volanté’). A possible explanation to this could be that we are faced with a composite work, partially preserving and partially rewriting and interpolating a work of Chrétien de Troyes (or at least mixing the styles of different authors).

To further explore the stylometric analysis of Chrétien authorship, and specifically the problems of *Guillaume d’Angleterre* and *Philomena*, further research should envision an authorship verification approach, such as ‘impostors’, with a larger corpus, including more contemporaries of Chrétien (for instance, Benoît de Sainte-Maure). In addition, the topic of the *Perceval* and all its *Continuations* should be investigated in more detail. Finally, other features that have proven interesting for medieval texts could be used in addition to lemmatised function words, in particular rhyme words, that are considered less subject to scribal variation (Kestemont et al., 2012).

Acknowledgments

This paper results from the collective work, made by the students of the MA *Computational Humanities* (Université PSL), in the Seminar of *Computational Philology* (year 2023-2024), led by Jean-Baptiste Camps, Florian Cafiero & Benedetta Salvati.

It benefited from funding from project *The Lost Manuscripts of Medieval Europe: Modelling the Transmission of Texts* (LostMA; ERC StG, 101117408). Student participation was funded by the École nationale des chartes, PSL.

The authors would also like to thank the four anonymous reviewers of DH Benelux 2024, for their useful suggestions.

References

- Emmanuèle Baumgartner, editor. *Philomena*. Paris, Gallimard, 2000. URL <http://catalog.bfm-corpus.org/philomena>.
- Matilda Tomaryn Bruckner. Authorial Relays: Continuing Chrétien’s Conte du Graal. In Virginie Greene, editor, *The Medieval Author in Medieval French Literature*, Studies in Arthurian and Courtly Cultures, pages 13–28. Palgrave Macmillan US, 2006. ISBN 978-1-4039-8345-9. doi: 10.1057/9781403983459_2. URL https://doi.org/10.1057/9781403983459_2.
- Florian Cafiero and Jean-Baptiste Camps. Why Molière most likely did write his plays. *Science advances*, 5(11):eaax5489, 2019.
- Florian Cafiero and Jean-Baptiste Camps. ‘Psyché’ as a Rosetta Stone? Assessing Collaborative Authorship in the French 17th Century Theatre. In Maud Ehrmann, Folgert Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas, Michael Piotrowski, and Joris van Zundert, editors, *Proceedings of the Conference on Computational Humanities Research 2021*, volume 2989 of *CEUR Workshop Proceedings*, pages 377–391, Amsterdam, the Netherlands, 2021. CEUR. URL http://ceur-ws.org/Vol-2989/#long_paper51. ISSN: 1613-0073.
- Jean-Baptiste Camps. SUPERvised STYLometry (SuperStyl), 2021. URL <https://github.com/SupervisedStylometry/SuperStyl>. original-date: 2021-02-26T13:38:37Z.
- Jean-Baptiste Camps, Thibault Clérice, Ariane Pinche, Lucence Ing, Frédéric Duval, and Naomi Kanaoka. Deucalion, Modèle Ancien Français (0.3.0), December 2020. URL <https://doi.org/10.5281/zenodo.4898954>.

- Jean-Baptiste Camps, Thibault Clérice, Frédéric Duval, Lucence Ing, Naomi Kanaoka, and Ariane Pinche. Corpus and Models for Lemmatisation and POS-tagging of Old French. *arXiv:2109.11442 [cs]*, 2021a. URL <http://arxiv.org/abs/2109.11442>.
- Jean-Baptiste Camps, Thibault Clérice, and Ariane Pinche. Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer’s hagiographic hypothesis. *Digital Scholarship in the Humanities*, 36:ii49–ii71, 2021b. ISSN 2055-7671. doi: 10.1093/llc/fqab033. URL <https://doi.org/10.1093/llc/fqab033>.
- Estelle Doudet. *Chrétien de Troyes*. Tallandier, Paris, 2009.
- Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32:ii4–ii16, 2017. ISSN 2055-7671. doi: 10.1093/llc/fqx023. URL https://academic.oup.com/dsh/article/32/suppl_2/ii4/3865676.
- Simon Gabay, Thibault Clérice, et al. Reconnaissance des écritures dans les imprimés. CATMuS print : un modèle générique, multilingue et diachronique, 2024. submitted to *Humanistica* 2024.
- Céline Guillot, Serge Heiden, and Alexei Lavrentiev. Base de français médiéval: une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques. Revue de Linguistique française diachronique*, 7:168–184, 2018. URL <https://shs.hal.science/halshs-01809581/>.
- Danièle James-Raoul. *Chrétien de Troyes, la griffe d’un style*. Honoré Champion, Paris, 2007.
- Mike Kestemont, Walter Daelemans, and Dominiek Sandra. Robust rhymes? the stability of authorial style in medieval narratives. *Journal of Quantitative Linguistics*, 19(1):54–76, 2012.
- Mike Kestemont, Sara Moens, and Jeroen Deploige. Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux. *Digital Scholarship in the Humanities*, 30(2):199–224, 2015.
- William W Kibler. L’Héritage de Chrétien de Troyes. *Cahiers de recherches médiévales et humanistes. Journal of medieval and humanistic studies*, 2007.
- Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. eScriptorium: An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–24, 2019. URL <https://ieeexplore.ieee.org/abstract/document/8893029/>.
- Roberta L. Krueger. *Philomena: Brutal Transitions and Courtly Transformations in Chrétien’s Old French Translation*, pages 87–102. Boydell & Brewer, Woodbridge, ned - new edition edition, 2005. ISBN 9781843840503. URL <http://www.jstor.org/stable/10.7722/j.ctt9qdhp.12>.
- Pierre Kunstmann, editor. *Chrétien de Troyes: Cligès, Erec, Lancelot, Perceval, Yoain – manuscrit P (BnF fr. 794)*. ATILF and Univ. Ottawa, 2009. URL <http://www.atilf.fr/dect>.

- Norris J. Lacy and Joan Tasker Grimbert, editors. *A Companion to Chrétien de Troyes*. Boydell & Brewer, Woodbridge, 2005. ISBN 978-1-84384-050-3. URL <https://www.jstor.org/stable/10.7722/j.ctt9qdhpb>.
- Félix Lecoy, editor. *Roman de la Rose ou de Guillaume de Dole*. Paris, Champion, 1962. URL <https://txm-bfm.huma-num.fr/txm/pdf/dole.pdf>.
- Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. Improving Lemmatization of Non-Standard Languages with Joint Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1493–1503. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1153. URL <https://aclanthology.org/N19-1153>.
- Francisque Michel, editor. *Roman de la violette, ou de Gérard de Nevers, en vers*. 1834. URL <https://archive.org/details/romandelaviolet00degoog/page/n58/mode/2up>.
- Petr Plecháč. Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns. *Digital Scholarship in the Humanities*, 36(2):430–438, 2021.
- Brian J Reilly and Moira R Dillon. Virtuous Circles of Authorship Attribution through Quantitative Analysis: Chrétien de Troyes’s Lancelot. *Digital Philology: A Journal of Medieval Cultures*, 2(1):60–85, 2013.
- Maria Slautina. The Disputed Authorship of a Medieval Text: A New Solution to the Attribution Problem in the Case of Guillaume D’Angleterre. *Authorship*, 2(1), 2012.
- Franklin Sweetser, editor. *Jean Renart: Escoufle*. Genève, Droz, 1974. URL <https://txm-bfm.huma-num.fr/txm/pdf/escoufle.pdf>.
- Karina Van Dalen-Oskam and Joris Van Zundert. Delta for middle Dutch—author and copyist distinction in Walewein. *Literary and Linguistic Computing*, 22(3):345–362, 2007.
- Philippe Walter. *Chrétien de Troyes*. Presses Universitaires de France, 1997.
- Harry F Williams. The Authorship of ‘Guillaume d’Angleterre’. *South Atlantic Review*, 52(1):17–24, 1987.
- Mary Williams, editor. *La Continuation de Perceval*. Robarts - University of Toronto, 1922. URL <https://archive.org/details/lacontinuationde00gerbuoft/page/vi/mode/2up>.
- Maurice Wilmotte, editor. *Guillaume d’Angleterre*. Honoré Champion, 1927. URL <https://gallica.bnf.fr/ark:/12148/bpt6k62267594>.
- François Zufferey. La pomme ou la plume: un argument de poids pour l’attribution de Guillaume d’Angleterre. *Revue de linguistique romane*, 285-286:157–208, 2008.

A. Corpus

Table 2: Composition of the corpus; all texts are Romances in octosyllabic verses (BFM = *Base de français médiéval*, Guillot et al. (2018); AO = *Archive.org*)

Author	Title	Genre	Date Work	MS Ed.	Date MS	Ed	Source
Chrétien de Troyes	Erec et Enide	Arthur.	1170	BNF, fr. 794 (C)	c. 1235	(Kunstmann, 2009)	BFM
Chrétien de Troyes	Cligès	Arthur.	1176	BNF, fr. 794 (C)	c. 1235	(Kunstmann, 2009)	BFM
Chrétien de Troyes	Yvain	Arthur.	1178-1181	BNF, fr. 794 (C)	c. 1235	(Kunstmann, 2009)	BFM
Chrétien de Troyes & Godefroi de Leigni	Lancelot	Arthur.	1178-1181	BNF, fr. 794 (C)	c. 1235	(Kunstmann, 2009)	BFM
Chrétien de Troyes	Perceval	Arthur.	1182-1190	BNF, fr. 794 (C)	c. 1235	(Kunstmann, 2009)	BFM
Chrétien 'li Gois'	Philomena	Antiqu.	1165-1170	Rouen O.4 (1044) (A1)	c. 1325	(Baumgartner, 2000)	BFM
Chrétien	Guillaume d'Angleterre	Advent.	12 ^{ex} -13 ⁱⁿ	BNF, fr. 375 (P)	1290-1317	(Willotte, 1927)	Gallica
Gerbert de Montreuil	Perceval's Fourth Continuation	Arthur.	1235	BNF fr. 12576 (A)	13 ^{3/4}	(Williams, 1922)	AO
Gerbert de Montreuil	Violette	Courtly	1227-1229	BNF fr. 1374+ fr. 1553	13 ⁱⁿ + 1285-90	(Michel, 1834)	AO
Jean Renart	Rose/Guillaume de Dole	Courtly	1200-11	BAV, Reg. lat. 1725	13 ^{ex} -14 ⁱⁿ	(Lecoy, 1962)	BFM
Jean Renart	L'Escoufle	Advent.	c. 1200-02	Arsenal 6565	13 ^{ex}	(Sweetser, 1974)	BFM

B. Coefficients of the SVM models

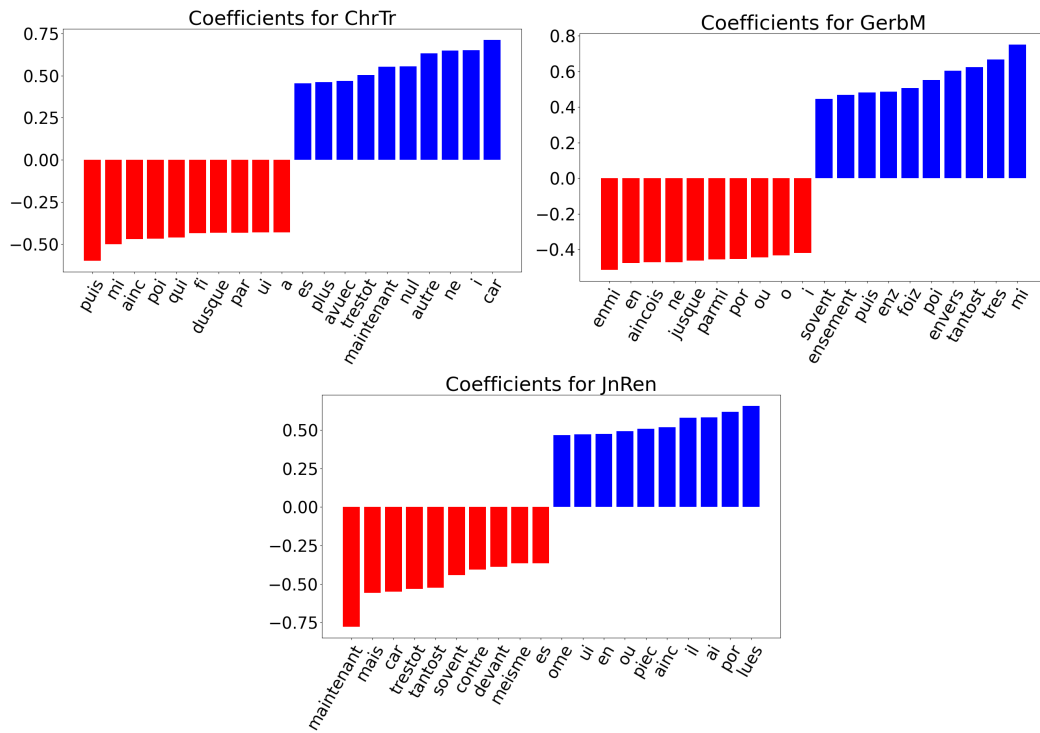


Figure 3: Coefficients of the SVM models for the three authors