Style in Eight Syllables: Metric Annotation and Stylometry of Chrétien de Troyes and Contemporaries

Florian Cafiero[1,2], Philippe Chaumet-Riffaud[1], Damien Conceicao[1], Ulysse Godreau[1], Emilie Guidi[1], Alexandre Lionnet[1], Théo Moins[1], Pierre-Alexandre Nistor[1], Benedetta Salvati[1], and Jean-Baptiste Camps[1]

[1]École nationale des chartes - PSL, Paris, France
[2]Institut des Hautes Etudes Internationales, Genève, Suisse

June 5, 2025

Authorship attribution for medieval texts such as those of Chrétien de Troyes poses unique challenges due to textual transmission, language variation, and limited reference corpora. In this context, it might be useful to draw as much stylistic information as possible from the texts, beyond most common features such as function words. This paper presents an ongoing project to include metrical annotation (with a focus on prosody) in the stylometric analysis of Medieval French, to enhance support vector machine (SVM)-based authorship attribution. The case at hand focuses on the attribution of the works of Chrétien de Troyes and his contemporaries.

Introduction

The works attributed to Chrétien de Troyes, a cornerstone of medieval French literary heritage (Kibler, 2007), are central to the discussion of stylistic and linguistic analysis of Medieval texts. The numerous authorship problems surrounding his work (Camps et al., 2024) however raise numerous challenges for the analysis of a sometimes hard to delimitate corpus. Issues such as spelling variation, scribal transmission, and incomplete corpora complicate the task, as well as the scarcity of data for some candidate authors. While standard computational methods such as Support Vector Machines (SVM) have been successfully employed in modern authorship attribution , they require reliable and consistent feature sets.

Recent research highlights the importance of metrical features as a valuable complement to traditional stylometric approaches in authorship attribution. Metrical analysis has proven particularly effective in poetic texts and structured prose, where rhythm and syllabic patterns can serve as distinct stylistic markers.

Corbara *et al.* 2023 demonstrate the utility of syllabic quantity in Latin prose, revealing that rhythmic preferences specific to authors significantly improve attribution accuracy when combined with machine learning models like Support Vector Machines (SVM). Application to non-literary texts (Corbara et al., 2022), such as Spanish parliamentary discourse, illustrate even further the effectiveness on using rythmic information authorship profiling. Similarly, Plecháč and Birnbaum 2023 show that rhythm-based features in poetry, particularly those deviating from strict accentual regularity, enhance attribution by highlighting unique authorial patterns. Across multiple languages and genres, metrical features have been shown to rival conventional lexical markers. Plecháč *et al.* 2018 confirm that versification analysis performs comparably to word frequency and n-gram-based methods in Czech, German, Spanish, and English poetry. This reinforces the cross-linguistic applicability of metrical features in stylometry. Applications to literary studies have already been succesful, as a recent study, combining the use of rhythmic and lexical analysis (Plecháč, 2021) to Shakespeare and Fletcher's collaborative work, has proven.

Material and Methods

To address the attribution questions, we selected a corpus of medieval French texts attributed to Chrétien de Troyes and other contemporary authors.

Corpus

The corpus was constructed around the works of Chrétien de Troyes (fl. c. 1170-1190), encompassing all his definitively attributed romances as well as the *Lancelot*, which he initiated and was presumably completed by Godefroi de Leigni (Adams, 1974). The disputed texts *Philomena* and *Guillaume d'Angleterre* were also included.

It is difficult to find relevant impostors from the same time period with comparable corpus of similar genres and for which we have enough confidence about authorship. Only two impostors were thus incorporated into the corpus: Gerbert de Montreuil (fl. first half of the 13th century) was selected as he is the author of a *Continuation* to Chrétien's works, thereby contributing texts closely aligned in genre, form, and content. Jean Renart (fl. end of the 12th-beginning of the 13th century) was included as an external reference.

To assemble the corpus, data were gathered from multiple sources. Chrétien's works utilized the digital edition of the copy by the scribe *Guiot de Provins* (Kunstmann, 2009). Other texts were sourced from the *Base de français médiéval* (Guillot et al., 2018), which provides digitized editions. For texts without available digital editions, OCR and post-correction were performed using eScriptorium (Kiessling et al., 2019) and the CATMUS Print model (Gabay et al., 2024).

Metric annotation

All the texts use the narrative octosyllabic couplet, typical of verse romances (and of most Old French narrative texts in verses, with the notable exclusion of *chansons de geste*), where 8 syllables verses are rhymed two by two, and end with either a masculine or feminine rhyme, with no standard caesura placement.

Metric annotation follows the 'Metronome' model, where **w** is used for weak syllables, **S** for stressed syllables, the dot (**.**) for word breaks, and | for verse separation (Nagy et al., 2024). The dot word separator was eliminated for further analysis, as

Verse	Annotation
et li chevax tant s'esforça	w.w.wS.w.wwS
Et a la pierre de l'anel,	w.w.w.Sw.w.wS

Table 1: Exemples of human annotation

preliminary tests showed it decreased accuracy of the SVM models, something already noted by Nagy et al. (2024). The question of the place of stress in French verse is a delicate and sometimes controversial one (see for instance Pensom, 2006). To facilitate this task, and obtain consistent results between annotators, two simple principles were followed:

1. function words are considered unstressed;
2. content words are stressed on the tonic syllable, i.e. the last syllable with the exclusion of final atone *-e* (*dame* is coded 'Sw', while *mangier* is coded 'wS').

An example is given in Table 1.

Metric Annotator: rule based vs LLM

The metric annotator developed in this study is designed to analyze the stressed syllables in Old French. To design it, we tried two approaches: relying on a generative AI annotator or designing a hybrid approach, using deep learning based linguistic annotation of parts-of-speech, complemented by a rule-based annotation of stress, relying on expert knowledge of Old French phonology. The task at hand is divided in both cases in two main steps: syllabification, then detection of stressed and unstressed syllables.

In order to better detect errors and to allow for an interim evaluation of the LLM, we decided to rely on a chain-of-thought approach to separate the syllabification task from the stress annotation task (Wei et al., 2022). A set of one hundred verses was initially annotated by hand as exemplars for the prompts, complemented with a description of the main rules to be followed for syllabification and stress annotation. The efficacy of the prompts was then tested in several LLMs. The results were disappointing, with an accuracy of 78.27% using Claude 3.5 Sonnet (the accuracy is computed as the average of the Levenshtein ratio between pairs of verses in the ground truth and the annotation).

In view of the unsatisfactory outcomes for the annotation obtained with the LLM, the decision was taken to eschew reliance on a generative AI annotator. First, a set of rules based on phonological knowledge of Old French (and particularly, its large number of diphthongs and triphthongs) is applied to divide verses into syllables as shown in appendix **??**. Then, the lemmatiser and POS-tagger from Deucalion model (Camps et al., 2020) are employed to facilitate the identification of function words and consequently weak syllables. Words such as adverbs with a DEGREE attribute in morpho-syntactic indications, verbs with the lemmas 'estre1' and 'avoir,' pronouns, and prepositions are considered function words. The rule-based approaches reaches an accuracy of 97.87% on the test set.

Table 2: Results using the different features. Optimal parameters for each configuration were found using a grid search (see Appendix). The two last scores correspond to a merge of two sets of features in order to be used in combination.

Feats	n	culling	F1 score
funct. words	1	0.1	0.86
met. syll.	11	0.6	0.88
met. lines	1	0.3	0.75
funct. words	1	0.6	0.95
+ met. syll.	9	0.7	
funct. words	1	0.4	**0.97**
+ met. lines.	1	0.9	

SVM-based Attribution Framework

We adopt a rolling stylometry (Eder, 2016) framework based on support vector machines (SVM), as implemented in the SuperSTYL Python package (Camps and Cafiero, 2024). The features selected are:

function words (baseline) lemmatised function words. Function words are classically used as features in authorship attribution (Mosteller and Wallace, 1963), as they are reasonably independent from the topics evoked in the texts, chosen less consciously, and in a sufficient numbers to yield statistically relevant results (Kestemont, 2014). Lemmatisation is sometimes avoidable for medieval texts when the corpus is linguistically very homogenous (Camps and Cafiero, 2013), which is unfortunately not the case here (see appendix "Corpus"). As in previous literature (Camps et al., 2021a), we thus applied lemmatisation to normalize the various spellings of Old French. Lemmatisation was conducted using Pie (Manjavacas et al., 2019) in conjunction with a model specifically trained for Old French (Camps et al., 2020, 2021b) Subsequently, function words were extracted from the lemmatised text using a custom list developed from the lemmatised Gold data of the 0F3C Corpus (Camps et al., 2021b).

rhythmic patterns n-grams of syllables, and of verses, from the metric annotation.

We systematically benchmark different length of n-grams and different levels of culling to find the best configurations (Table 2). Results show that prosodic features, such as n-grams of stressed and unstressed syllables, perform comparably to the baseline in isolation. They seem to bring sufficient information by themselves for reliable attribution of the texts in the train set. Additionnally, when taken in combination with function words, they allow to substantially increase the reliability of the attribution.

Results

Results obtained with the two combined setups (n-grams of stressed syllables with function words; 1-grams of prosodic annotation of verses with function words) seem to confirm the drop in Chrétien de Troyes's style around verse 6150 in *Lancelot*, once again supporting the attribution of the final part to another author, i.e. Godefroi de Leigny (Figure 1). Intriguingly, a similar drop is observed around the verse 2000, which could hint at the intervention of a second hand in other places of the romance.

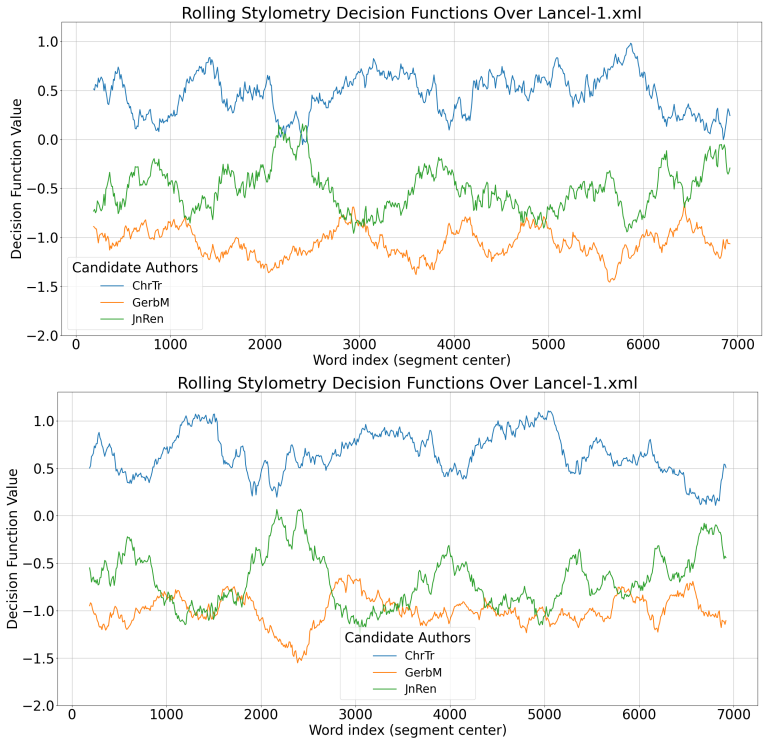

Figure 1: Rolling analysis of *Lancelot*, using function words and stress annotation of syllables (up), or function words and stress annotation of full verses (down)

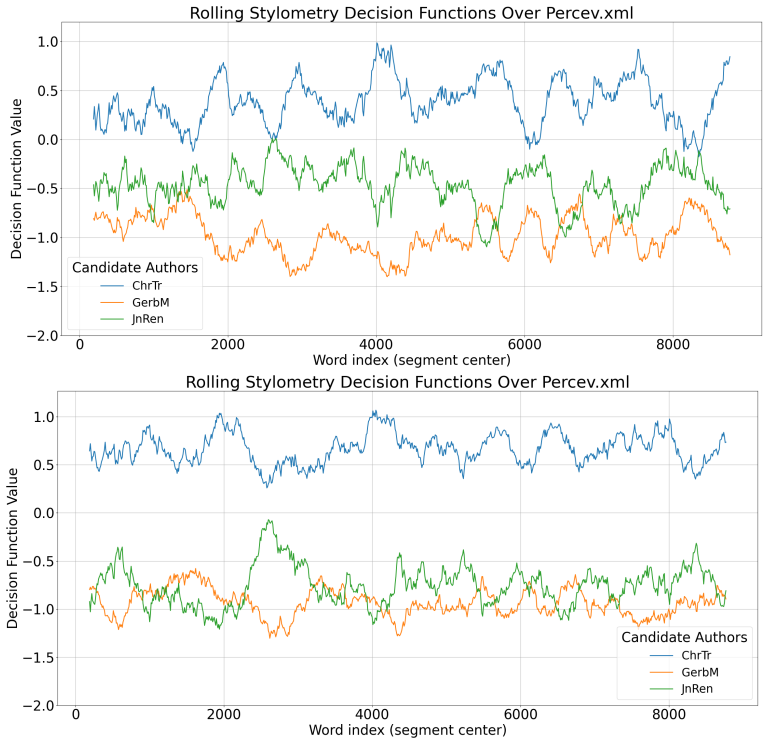

Figure 2: Rolling analysis of *Perceval*, using function words and stress annotation of syllables (up), or function words and stress annotation of full verses (down)

On the other hand, concerning *Perceval*, the results are more mixed. The first configuration shows a drop in Chrétien's style towards the end, already observed in a previous study (Camps et al., 2024), while the second (and best) configuration makes this variation disappear and seem to confirm the stylistic homogeneity of the text (Figure 2).

Discussion

In the current state of research, our analyses show that the annotation of stressed and unstressed syllables seem to provide reliable features for authorship attribution of medieval French verse texts. Alone, they perform reasonably well, and are at their best when used in combination with other traditional features, such as function words, bringing additional relevant stylistic information, something already observed in other contexts (Plecháč et al., 2018).

Regarding the case study, the inclusion of these new features improve the reliability of the analyses, and confirm the collaborative nature of *Lancelot*. On the other hand, the situation of *Perceval* is still deserving further investigation. As for now, the instability of the drop of Chrétien's style at the end, from one analysis to the other, would lead us to discard the stronger hypothesis of a collaborative nature of this text, not attested through external evidence.

Code and data availability

The code of the prosodic annotator is available on Github, `https://github.com/PoidsPlume/AnoChre`.

The rest of the data and code is available on Github, `https://github.com/Humanites-Numeriques-PSL/chrestien/` and released on Zenodo.

Acknowledgments

This paper results from the collective work, made by the students of the MA *Digital Humanities* (Université PSL) in the Seminar of *Computational Philology* (years 2023-2025), led by Jean-Baptiste Camps, Florian Cafiero & Benedetta Salvati.

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

References

Alison Adams. Godefroi de leigni's continuation of lancelot. In *Forum for Modern Language Studies*, volume 10, pages 295–299. Oxford University Press, 1974.

Jean-Baptiste Camps and Florian Cafiero. Setting bounds in a homogeneous corpus: a methodological study applied to medieval literature. *Revue des Nouvelles Technologies de l'Information*, (MASHS 2011/2012. Modèles et Apprentissages en Sciences Humaines et Sociales Rédacteurs invités: Mar):55–84, 2013.

Jean-Baptiste Camps and Florian Cafiero. SUPERvised STYLometry (SuperStyl), 11 2024. URL https://doi.org/10.5281/zenodo.14069799.

Jean-Baptiste Camps, Thibault Clérice, Ariane Pinche, Lucence Ing, Frédéric Duval, and Naomi Kanaoka. Deucalion, Modèle Ancien Francais (0.3.0), December 2020. URL https://doi.org/10.5281/zenodo.4898954.

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(Supplement_2):ii49–ii71, 2021a.

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*], 2021b. URL http://arxiv.org/abs/2109.11442.

Jean-Baptiste Camps, Benedetta Salvati, Gonzalo Freijedo, Donghan Bian, Gaëtan Drouet, Eglantine Gaglione, Émilie Guidi, Carolina Macedo, Yaelle Zribi, and Florian Cafiero. The authorship of the works of chrétien de troyes: a stylometric examination. In *DH Benelux 2024*, 2024.

Silvia Corbara, Berta Chulvi, Paolo Rosso, and Alejandro Moreo. Rhythmic and psycholinguistic features for authorship tasks in the spanish parliament: Evaluation and analysis. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 79–92. Springer, 2022.

Silvia Corbara, Alejandro Moreo, and Fabrizio Sebastiani. Syllabic quantity patterns as rhythmic features for latin authorship attribution. *Journal of the Association for Information Science and Technology*, 74(1):128–141, 2023.

Maciej Eder. Rolling stylometry. *Digital Scholarship in the Humanities*, 31(3):457–469, 2016.

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/.

Mike Kestemont. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, 2014.

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/.

Pierre Kunstmann, editor. *Chrétien de Troyes: Cligès, Erec, Lancelot, Perceval, Yvain – manuscrit P (BnF fr. 794)*. ATILF and Univ. Ottawa, 2009. URL `http://www.atilf.fr/dect`.

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`.

Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302): 275–309, 1963.

Ben Nagy, Artjoms Šeļa, Mirella De Sisto, and Petr Plecháč. Metronome: tracing variation in poetic meters via local sequence alignment. *arXiv preprint arXiv:2404.17337*, 2024.

Roger Pensom. Pour la versification anglo-normande. *Romania*, 124(493/494 (1/2): 50–65, 2006.

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.

Petr Plecháč and David J Birnbaum. Rhythm-based authorship recognition in syllabic and accentual-syllabic verse. *Literatura: teoría, historia, crítica*, 25(2), 2023.

Petr Plecháč, Klemens Bobenhausen, and Benjamin Hammerich. Versification and authorship attribution. a pilot study on czech, german, spanish, and english poetry. *Studia Metrica et Poetica*, 5(2):29–54, 2018.

Jason Wei et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

A. Prosodic annotation: rule-based approach

Each verse is processed letter by letter using the following rules:

- Basic Rules:
  - Automatic addition of consonants to the current syllable.
  - Addition of vowel followed by the termination of the syllable.
- Complex Rules:
  - If the last syllable of a word contains only consonants, it is attached to the preceding syllable.
  - If the analyzed vowel and the following vowel form a diphthong that does not end in 'e', both are added to the current syllable.

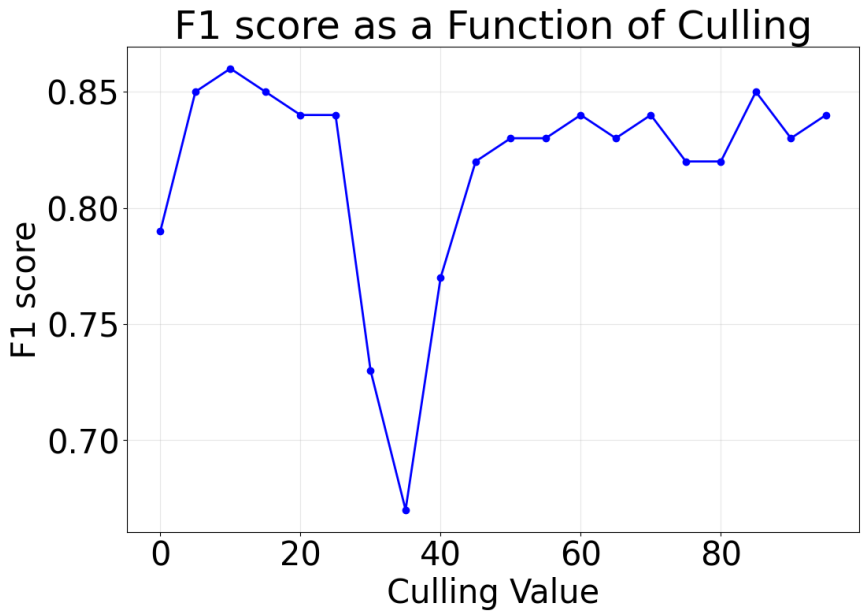

Figure 3: **Left**: Evolution of the f1-score for the SVM using function features with different values of culling.

- If the analyzed vowel and the following vowel form a diphthong that ends in 'e', it is checked whether the 'e' is at the end of the word; if so, it is separated into its own syllable.
- Unknown characters are removed.
- Apostrophes are treated as consonants

By default, all syllables are marked as weak. Then, if the part-of-speech of the word is not inside the function words categories, the last syllable is marked as stress, except if it ends in *-e*. If it ends in *-e*, then the penultimate syllable is marked as stressed.

B. Benchmarking function words and prosodic features

Here, we present our iterations to obtain the best f1-score for each combination of features we consider. For each iteration, we create a dataset based on the features characteristics, and then we train a SVM with group K-fold cross validation. In particular, two parameters were studied here: the n-gram size n, and the ratio of culling.

B.1. Attribution using only function words

We iterate along different values of culling from to 0 to 90% with a step of 5% for functions words. Results are shown in Figure 3 and suggests a value of 10%.

B.2. Attribution using only prosodic features

For the case of stress annotated syllables, we iterate on values of n and culling, and for entire verses, only culling is iterated for 1-gram and 2-gram of verses. We only show the results of 1-gram here, as using 2-gram of verses was not satisfying. Results are shown in Figure 4.

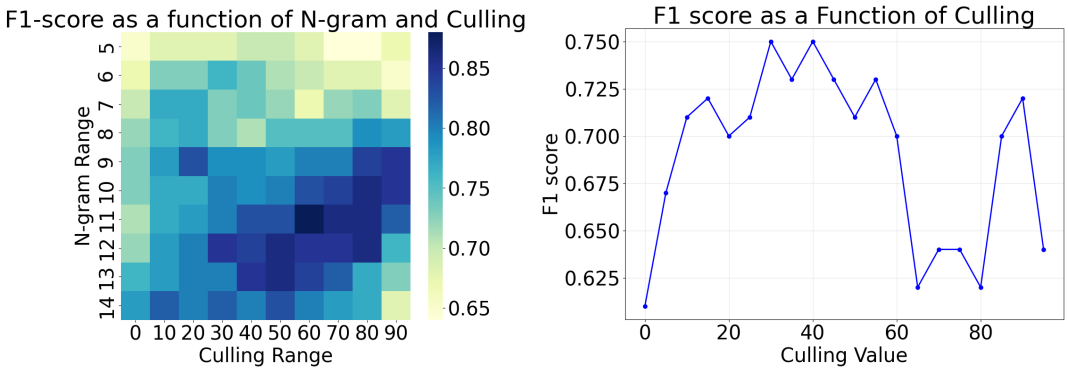

Figure 4: **Left**: Heatmap of the f1-score for the SVM using stress annotated syllables with different values of n and culling. **Right**: heatmap of the f1-score for the SVM using 1-gram of verses with different values of culling.

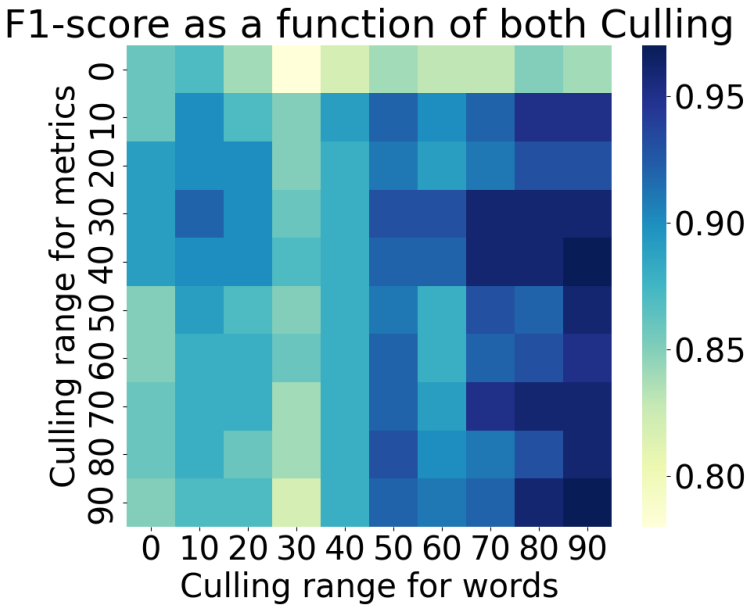

Figure 5: Heatmap of the f1-score for the SVM using function words and stress annotated verses with different values of culling for both type of features.

B.3. Combining the best of both worlds

For function words with stress annotated syllables, our grid search is 3-dimensional: n-gram length of function words, culling ratio of function words, and culling ratio of syllables. In the combination of function words and full verses, we focus on 1-gram of verses and vary the culling range for both function words and verses (see Figure 5).

C. Suplementary rolling analyses

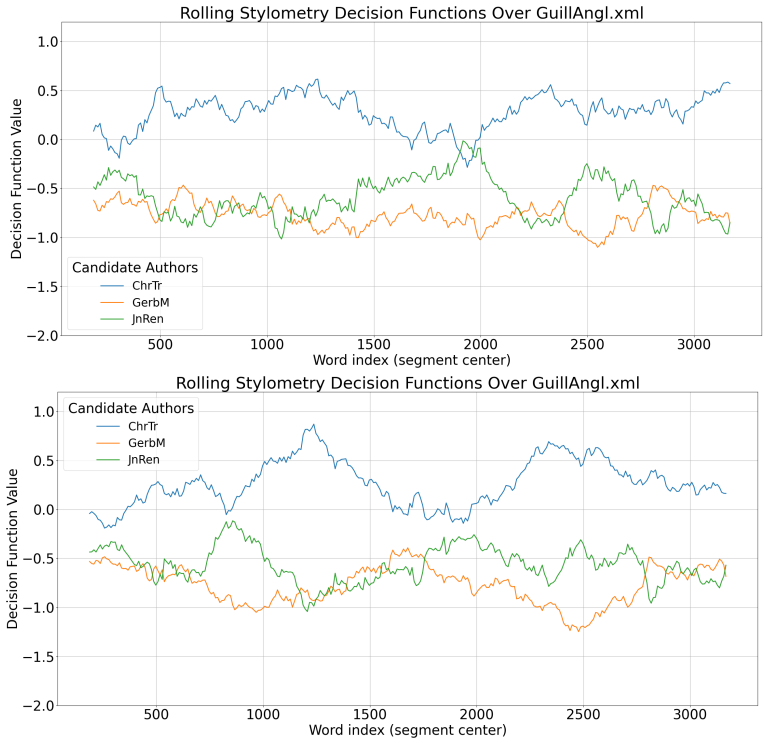

Figure 6: Rolling analysis of *Guillaume d'Angleterre*, using function words and stress annotated syllables (up), or function words and stress annotated verses (down)

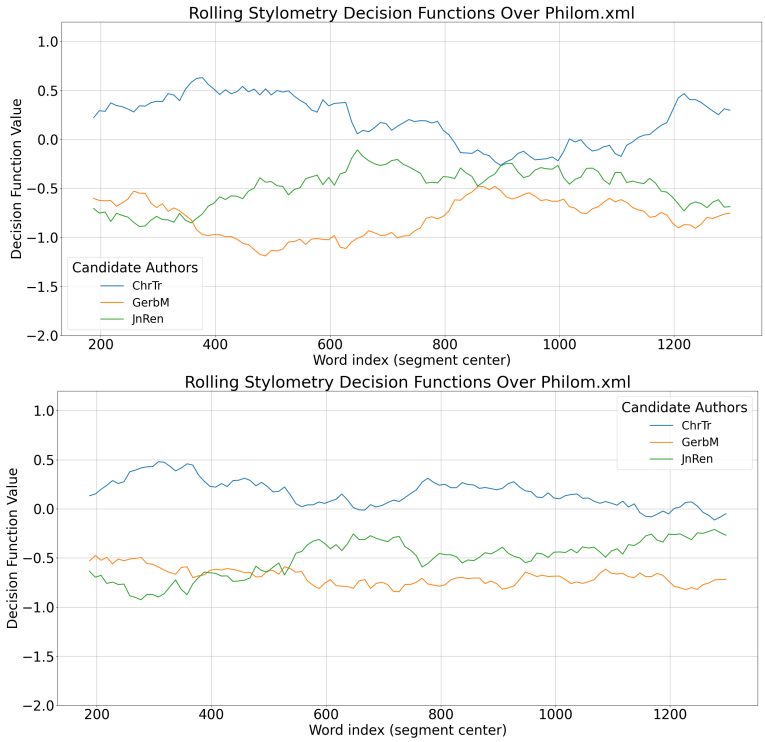

Figure 7: Rolling analysis of *Philomena*, using function words and stress annotated syllables (up), or function words and stress annotated verses (down)

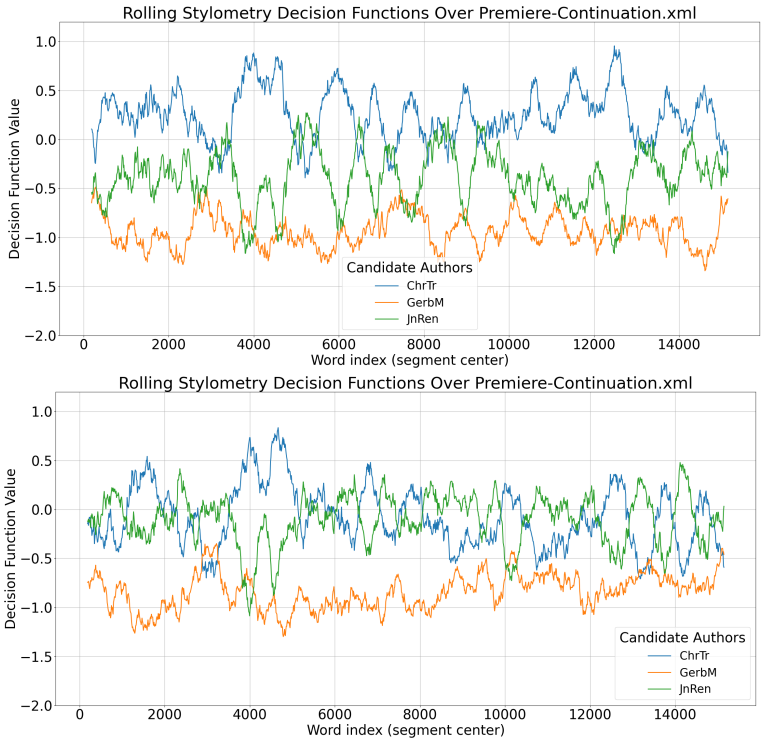

Figure 8: Rolling analysis of *Première Continuation*, using function words and stress annotated syllables (up), or function words and stress annotated verses (down)

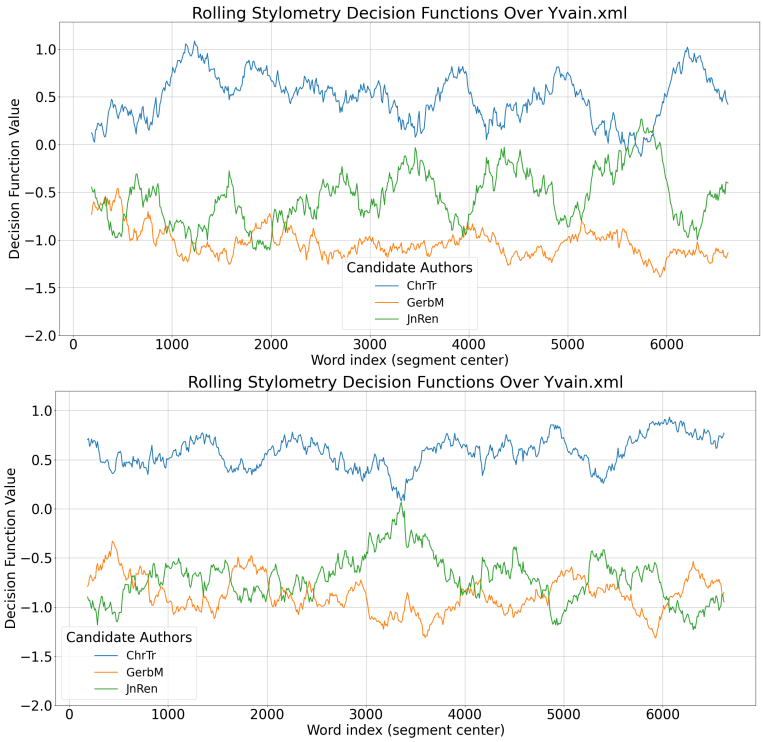

Figure 9: Rolling analysis of *Yvain*, using function words and stress annotated syllables (up), or function words and stress annotated verses (down)