Echoes of the Troubadours
A Corpus of Troubadour Poetry for Stylometric Analysis and Authorship Attribution

Loic De Langhe, Orphée De Clercq, Veronique Hoste
LT[3], Language and Translation Technology Team, Ghent University, Belgium
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

We present TrovaCor, a curated corpus of medieval troubadour poetry, which comprises 1668 unique Old Occitan texts by a large variety of authors. Clustering and stylometric experiments show that we can accurately model authorial style beyond topical content, even though formulaic or topically diverse genres remain challenging. Furthermore, we can model and detect traces of an author's stylistic traits even in short-form collaborative poetry, offering a uniquely fine-grained perspective in the field. In addition, we provide self-organizing map visualizations in order to provide an interpretable view of stylistic patterns across authors. TrovaCor is publicly released to support reproducible research in NLP and digital humanities on this low-resource historical corpus.

Keywords: Medieval Occitan, Digital Humanities, Stylometry

1. Introduction

The troubadours (fem: Trobairitz) were a band of traveling poets and singers who roamed the courts of medieval Europe between the 11th and 13th centuries. Among their ranks were royals and local nobles, but also wandering knights and convicted criminals. The epicenter of their activity lay in the courts of medieval Occitania, a culturally rich region that today spans southern France as well as parts of Spain, and Italy. They composed in Old Occitan, a now-extinct language consisting of mutually intelligible dialects without standardized writing system (Paden, 1998). While orthographic practices varied considerably across regions and manuscripts, a fairly uniform poetic koiné emerged, which allowed troubadour verse to circulate across courts in medieval Europe (Field, 2006).

Today, the lyrical poetry of the troubadours is most often identified with the ideals of courtly love (fr: *L'amour Courtois*) and chivalry. While it is indeed the case that the majority of an estimated 2,500 surviving troubadour poems are classified as love songs, there exists a plethora of highly varied styles and genres within the tradition. These range from prayers to sarcastic poems mocking contemporary events, cryptic riddles, insults and highly explicit and even obscene verses. Consider the following excerpt by the 12th century singer Uc de Mataplana, directed at an individual known as Reculaire, as a typical example of troubadour style and mud-slinging:

Scometre·us vòlh, Reculaire;
Pois vestirs no·us dura gaire,
De paubretat ètz confraire
Als bons òmes de Leun;
Mas de fe no·n semblatz un
Que vos ètz fòls e jugaire,
E de putans cortejaire.

Translation: I challenge you, Reculaire; Since your clothes won't last long, you are a brother in poverty. To the good people of Lyon. But you don't seem like one in belief for you are a fool and a gambler, and a suitor of whores.

The troubadour body of work offers an interesting opportunity to examine a topically diverse and stylistically varied art form composed in the vernacular of the medieval period. Unlike the formal Latin commonly used in medieval literature, these texts provide insights into the everyday language and cultural expression of that time. In this paper, we bring together a diverse selection of troubadour texts into a single open-source digital corpus. By making this resource freely available, we aim to support the study of this fascinating, non-standard medieval language in both the Natural Language Processing (NLP) and Digital Humanities (DH) communities. Beyond compiling the corpus, we also demonstrate the potential of using modern computational methods through several experiments on (unsupervised) authorship attribution.

Our specific contributions are as follows:

- Release of the **TrovaCor dataset, comprising 1,668 original troubadour poems**, making it, to our knowledge, the largest online repository of medieval troubadour texts.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 905–918
11–16 May 2026.

- All **texts are cleaned, sourced, and annotated by genre**. Collaborative poems and references to other troubadours in the texts are **fully mapped**, creating a detailed network of author interactions.
- Corpus- and document-level authorship attribution experiments, which is the **first large-scale computational study of its kind on Old Occitan**.

2. Related Work

2.1. Old Occitan

Old Occitan, often referred to as Old Provençal, was a Romance language spoken across much of southern France and extending into parts of northern Spain and northwestern Italy between the 9th and 14th centuries (Paden, 1998). Typologically, Old Occitan shares several features with its Romance neighbors: for instance, it preserves a relatively conservative vowel system, shows early development of articles, and displays morphological characteristics closer to Catalan than to French (Sauzet, 2012; Wolfe, 2018). The language was not monolithic but encompassed a spectrum of regional dialects, each with distinct phonological and lexical traits (Sibille, 2024).

Figure 1: The approximate area of Medieval Occitania, overlaid with modern departmental and national borders

The bulk of the currently available written material in Old Occitan is lyric poetry. Most surviving troubadour poems, including a smaller amount of accompanying melodies, were preserved in songbooks or *Chansoniers*, richly illustrated compilations often made for wealthy patrons. During the revival of interest at the turn of the 20th century, many of these poems were arduously transcribed into critical editions (Appel, 1892; d'Ussel, 1922; Bec, 1984). Early attempts to categorize Old Occitan manuscripts were made by Bartsch (1872), who introduced a lettering system for identifying original sources. This scheme remains in use today, even in the digital age, where online relational databases such as those maintained by Brigham Young University [1] and the Sapienza University of Rome (Asperti and Nigro, 2003) provide access to a subset of the material. Other than poetry, Old Occitan also appears marginally in legal charters, didactic works, and religious texts, offering a valuable, if uneven, glimpse into its variation and historical development (Shields, 1976; Wilson, 2012).

NLP studies on Old Occitan remain extremely limited, largely due to the scarcity of available data. A major bottleneck is that most large relational databases, such as those mentioned earlier, provide only excerpts rather than complete texts. In recent years, initiatives like the Corpus des Troubadours (d'Estudis Catalans, 2020), Rialto (Di Girolamo et al., 2012), and the Dictionnaire de l'Occitan Médiéval (DOM) (Stempel et al., 1996) have made valuable contributions by offering digital resources, but such efforts are scarce. Despite the challenging nature of the task, recent years have seen an increased interest and progress has been made in both resource creation and language processing for Old Occitan. For prose some success was achieved through methods such as cross-lingual transfer in the context of constructing treebanks (Scrivner and Kübler, 2012). Beyond the digital corpora, Handwritten Text Recognition (HTR) has shown promise for building datasets (Arias et al., 2023), and more recently, Schöffel et al. (2025) presented a comprehensive evaluation of modern LLMs for POS tagging Old Occitan prose. To date, however, none of these studies have focused on the lyrical poetry of the troubadours.

2.2. Historical Authorship Modeling

Computational authorship attribution (AA) and stylometry aim to identify the author of a text and characterize writing style through measurable linguistic patterns. Stylometry broadly studies these features to describe and compare texts, while AA applies them to determine the likely author of an anonymous or disputed work. Historical authorship modeling extends these approaches across time, allowing scholars to trace the evolution of an author's style, detect diachronic linguistic shifts, and contextualize writing practices within broader cultural and scribal trends (De Gussem et al., 2022; Vandyck et al., 2025). In medieval studies, these methods are particularly valuable for analyzing anonymous or collaboratively produced works, helping to uncover authorial fingerprints and assess questions of attribution across manuscripts and poetic corpora. They also provide empirical grounding for

[1] https://troubadours.byu.edu/indextro.htm

attribution debates that have traditionally relied on subjective stylistic judgments. (Stamatatos, 2009; Jafariakinabad, 2021).

Interest in computational AA dates back to the origins of Computational Linguistics (Holmes, 1998), with early work often focusing on contested English texts such as the *Federalist Papers* (Mosteller and Wallace, 1963; Tweedie et al., 1996). More recently, AA research has expanded to a wider range of languages, including Dutch (Kestemont, 2012; Morante et al., 2022), Ancient Greek (Gorman and Gorman, 2016), and Spanish (López-Escobedo et al., 2013) (Savoy, 2020).

Stylometric studies have, among others, addressed authors like Hildegard of Bingen (Kestemont et al., 2015), Dante (Corbara et al., 2019), and Apuleius (Stover et al., 2016). Most of these studies rely on handcrafted stylistic features with traditional machine learning (Corbara et al., 2023).

As far as methodology is concerned, where early work used supervised learning with gold-standard labels, unsupervised methods are increasingly applied to uncover implicit similarities or outliers without prior annotation (Kehler and Stolcke, 1999; Bharadiya, 2023; Martín-del Campo-Rodríguez et al., 2022). Common techniques include agglomerative clustering (Layton et al., 2013; Panicheva et al., 2019), c-means (Demir, 2013), and self-organizing maps (Ranatunga et al., 2011; Neme et al., 2015), often leveraging easily extractable features such as character n-grams, punctuation, or simple text similarity measures (Kapočiūtė-Dzikienė et al., 2015; Tanguy et al., 2012; Qian et al., 2015). More recent work, has taken advantage of transformer-based NLP and has shown that deep semantic embeddings can be used to distill an author's stylistic signature (De Langhe et al., 2024). The advent of these approaches in medieval stylometry eliminates the need for extensive feature engineering while still achieving strong results in unsupervised settings.

3. Corpus Creation

3.1. Data Collection and Considerations

Most of the material in the original Occitan songbooks has been transcribed and preserved in critical editions, which will form the basis of our own data collection. Note that, despite these preservation efforts, a not insignificant part of the available songs are incomplete, anonymous, or of uncertain authorship (Klingebiel, 1997), and often multiple transcriptions of the same poem exist depending on the particular *Chansonier*. To ensure quality and consistency, we set two prerequisites for including a poem in the corpus. First, the source of the poem's transcription must be known and will be appended to the corpus metadata. Second, for poems with uncertain or disputed authorship, all possible attributions will be recorded in the metadata.

Data for the corpus was collected from several established online repositories such as the Corpus des Troubadours (d'Estudis Catalans, 2020) and the Rialto Project (Di Girolamo et al., 2012). After extracting the (Old) Occitan sections from these collections, the texts were manually verified, and their sources, possible attributions, and references to other troubadours within the texts were documented in the metadata. In addition to poems collected from these online sources, a significant number of texts were also transcribed from critical and diplomatic volumes on troubadour literature available through initiatives such as project GuthenBerg (Pro, 2025) and Galica (Gal, 2025). The procedure for transcription consisted of running an Optical Character Recognition algorithm through the scanned critical editions, followed by manual correction. For a detailed overview of the various sources from which the poems were drawn, we refer the reader to the corpus metadata [2]. In total, we collected 1668 lyrical poems, which can all be dated between the 11th and 14th century. These collection efforts result, to our knowledge, in the largest digital repository of (unique) troubadour poems in existence.

3.2. Corpus statistics

3.2.1. Genre

For each poem, we include its (presumed) author, the critical edition source, and its genre. As noted in the introduction, troubadour styles are highly varied, often distinguished by theme or stylistic patterns, and there is no fixed set of genres. While sources like the *Leys d'amors* define certain poem types and rules, changes over the centuries make it difficult to enumerate genres precisely (Bec, 1982; Chaillou, 2009). Genres are also fluid; a love song may take on a religious tone if dedicated to a saint, or an appraisal can become a satire when written with mocking intention.

Given these considerations, genre annotation in the corpus is not always clear-cut or particularly helpful. In order to mitigate this, we assign each poem two labels: a coarse-grained thematic label and a fine-grained label. The coarse-grained scheme groups poems into seven categories: Love songs, Satires, Religious songs, Collaborative songs, Short verses, Mourning songs, and Dancing songs – capturing the overall genre distribution. An overview of the entire corpus based

[2] https://github.com/Loicdelanghe/TrovaCor

on this taxonomy is shown in Table 1. The fine-grained label then corresponds to the genre as attested in the literature. At present, the corpus includes a total of 41 different fine-grained genres. A detailed table showing the distribution of fine-grained genres is provided in Appendix A, along with brief definitions for each in B for reference.

Category	Documents	Tokens
Love Songs	839	277,333
Satirical Songs	329	100,533
Collaborations	168	47,244
Short Verse	133	11,343
Religious Songs	98	32,005
Mourning Songs	35	14,318
Dancing Songs	25	4,583
Other	49	22,940
Total	1668	510,307

Table 1: Corpus statistics per coarse-grained category: number of documents and total (word) token counts.

3.2.2. Authorship

Works by a total of 347 individual authors are included in the corpus. The majority of these authors, 63% or 220 authors, have contributed only 3 or fewer poems to the corpus, while almost half of all songs were written by the 20 most prolific authors alone. While inconvenient, this imbalance is simply the result of some *Chansoniers* (or copies thereof) surviving longer than others. Going by the *vidas* of some of the less represented authors in the corpus, it is often the case that these poets were both renowned and industrious. However, more often than not, the majority of their oeuvre was unfortunately lost at some point in time.

A final consideration is the anonymous portion of the corpus: 95 unattributed poems, mostly short-form poetry or fragments of larger pieces. While these poems lack clear attribution leads, they nonetheless present interesting research opportunities for computational attribution methods and even generative reconstruction.

3.2.3. Collaborations

There exist several genres in the troubadour literature that typically have two (or more) authors, ranging from a short exchange of stanzas, to longer philosophical discussions. The text collection contains 168 of such collaborative poems, making up around a little over 10% of the entire corpus. Note here that the *Fictional Tenso*, in which the poet has a debate with a fictional (e.g. an idealized lover) or non-human (e.g. God, a horse) is also included in this number.

Interestingly, the interactions between various troubadours are not only limited to collaborations between them. Often times, the poets would explicitly mention both contemporary and past troubadours. These works can take the form of a *Plahn* (mourning song) or a *Sirventes* (Parody) in which the author either sings the praises of a deceased contemporary or insults them in subtle (or not so subtle) ways. Notorious examples of the latter include Uc de Mataplana's *Scometre·us vòlh, Reculaire* (**ENG:** *I challenge you, Reculaire*), which despite its formal title quickly results into name calling and some rather unsavoury accusations about his opponent's nightlife, as well as Peire Cardenal's *D'Esteve de Belmon m'enueia* (**ENG:** *To Esteve de Belmon, who annoys me*).

These examples highlight that troubadour works – and the data analyzed here – are best understood as a network of interactions between texts rather than isolated items. To illustrate this, Figure 3 presents a network of all collaborations and mentions among authors, with nodes representing authors and edges indicating collaborations or references. For clarity, only the most prominent authors are emphasized in a different color. The remaining (black) nodes refer to less prominent authors.

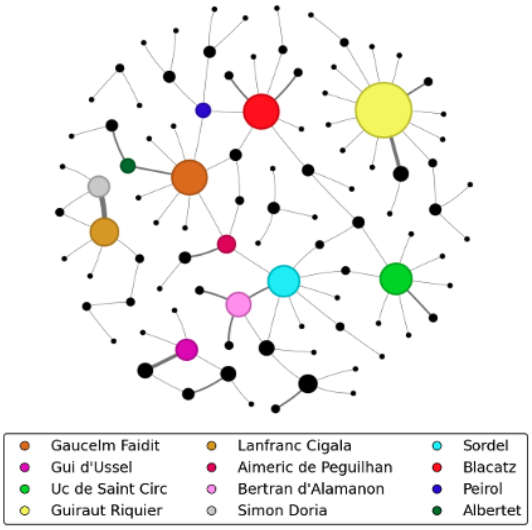

Figure 2: Visualisation of various interactions and collaborations between different troubadours.

4. Authorship Attribution

As a part of releasing this new medieval corpus, we also include a series of experiments and analyses for authorship attribution. Through these experiments we aim to illustrate to the reader a concrete example of how our dataset may be used within the broader domain of (computational) medieval studies.

		Intrinsic			Extrinsic		
		SC ↑	CHI ↑	DB ↓	RI ↑	ARI ↑	MR ↓
Monolingual	bert-base-italian-cased	0.10	66.39	2.41	0.87	0.35	0.49
	bert-base-spanish-wwm-uncased	0.05	49.67	3.12	0.85	0.24	0.54
	camembert-base	0.10	**67.5**	2.25	0.88	0.37	0.43
	deberta-small-occitan	0.10	33.72	2.21	0.86	0.31	0.46
	ModernBERT-base	0.04	18.46	3.80	0.80	0.05	0.72
	roberta-base-catalan	0.08	28.54	2.37	0.85	0.28	0.51
	roberta-base-latin-cased	0.06	31.87	2.78	0.68	0.04	0.76
Multilingual	bert-base-multilingual-cased	0.08	49.02	2.81	**0.89**	**0.43**	**0.4**
	Gemini text-embedding-001	0.01	8.51	4.21	0.79	-0.004	0.82
	mmBERT-base	0.00	0.00	0.00	0.00	0.00	0.00
	paraphrase-multilingual-MiniLM-L12-v2	0.01	7.35	4.30	0.83	0.16	0.60
	Qwen3-Embedding-0.6B	0.03	8.44	4.09	0.80	0.005	0.79
	xlm-roberta-base	0.09	42.14	**2.12**	0.85	0.23	0.54
	xlm-roberta-longformer-base-4096	0.08	33.4	2.51	0.86	0.27	0.46

Table 2: Clustering evaluation metrics with various monolingual and multilingual encoder models.

4.1. Corpus-level Authorship Attribution

Authorship attribution in medieval lyric is a challenging, fine-grained, multi-label task, as subtle stylistic variation and shared conventions complicate clear separation. In this experiment, we perform unsupervised clustering to evaluate our ability to capture these signals, focusing on the ten most prominent authors (450 works) and filtering out collaborative poems to reduce confounding factors. Through creating this subset, we reduce sparsity and can provide a clearer view of clustering performance.

4.1.1. Methodology

For each document d, we generate a vector representation using an encoder model. If the document exceeds the model's context length, it is split into chunks $c_1, c_2, \ldots, c_m$, and the embeddings of all chunks are averaged to produce a single document representation:

$$\mathbf{v}_d = \frac{1}{m} \sum_{k=1}^{m} \frac{1}{|T_k|} \sum_{t \in T_k} \mathbf{e}_t,$$

where T_k denotes the set of tokens in chunk c_k and $\mathbf{e}_t$ is the embedding of token t. Averaging over all tokens ensures that each part of the document contributes to the final representation.

We apply this procedure using a series of state-of-the-art monolingual and multilingual encoder models, allowing us to assess the robustness of clustering across different embedding architectures and languages. We specifically select monolingual encoders from typologically close languages such as Italian, Spanish, French, Modern Occitan, Catalan, Latin as well as the all-purpose English ModernBERT architecture (Warner et al., 2024). The mix of multilingual models consists of established BERT and RoBERTa architectures (Devlin et al., 2018; Liu et al., 2019) as well as three state-of-the-art multilingual embedding models in Gemini's `text-embedding-001`, the sentence encoder `multilingual-MiniLM-L12-v2` and a lightweight Qwen3 embedding model. Once all document embeddings are obtained, k-means clustering is applied to group documents based on embedding similarity.

Clustering performance is evaluated with both intrinsic and extrinsic metrics. Intrinsic metrics assess cluster quality without gold labels: both the Silhouette Coefficient (SC) and Calinski–Harabasz Index (CHI) favor compact, well-separated clusters, and the Davies–Bouldin Index (DB) penalizes overlap. Extrinsic metrics compare clusters to author labels: the Rand Index (RI) measures overall label agreement, the Adjusted Rand Index (ARI) corrects this for chance, and the Misclassification Rate (MR) reports the proportion of erroneously assigned documents.

4.1.2. Results and Analysis

Quantitative Results Table 2 highlights differences between monolingual and multilingual embeddings. Monolingual models like the French `camembert-base` achieve the best intrinsic scores, forming compact, well-separated clusters, while models trained on more distant languages, most notably English, perform worse. The multilingual models also perform best when evaluated through extrinsic metrics: `bert-base-multilingual-cased` achieves the highest Rand Index (RI = 0.89) and Adjusted Rand Index (ARI = 0.43), indicating a fair alignment with true author labels. Overall, monolingual models thus produce tighter clusters, whereas

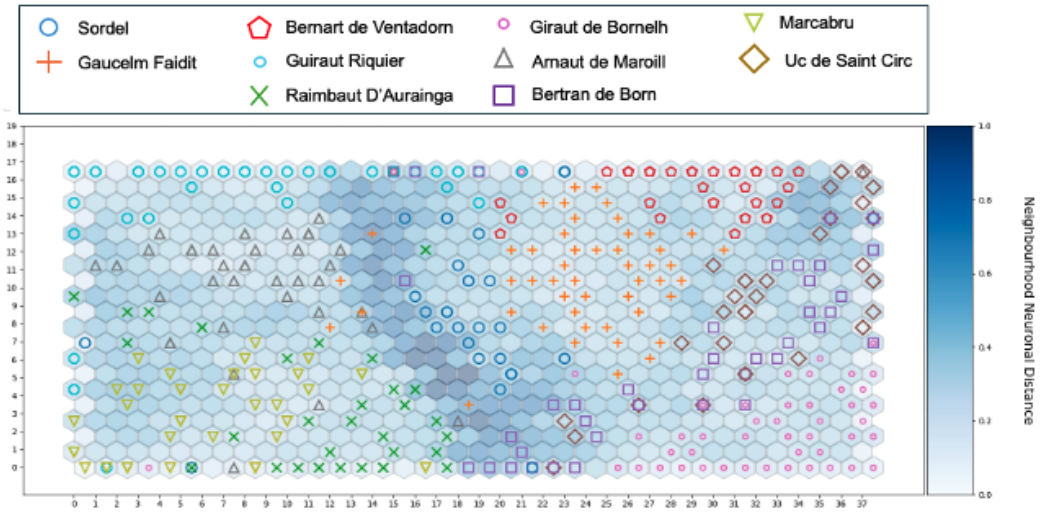

Figure 3: Visualisation of the trained self-organizing map using chunked mbert embeddings as document representations.

selected multilingual models better capture author-level distinctions reflected in the RI and ARI.

Qualitative results To complement quantitative evaluation, we use a self-organizing map (SOM) to produce an interpretable topographic visualization of document embeddings.[3] Following the standard algorithm (Oja and Kaski, 1999; Kohonen, 2013), high-dimensional embeddings are projected onto a two-dimensional grid that preserves relative distances. This provides an intuitive spatial layout for examining authorial similarity, stylistic overlap, and model performance.

The clustering results reveal well-separated author clusters in the topographical lattice, indicating that the embeddings effectively capture author-specific stylistic signals. Outliers are rare and usually correspond to works that differ topically from the majority of an author's corpus (e.g. a Crusader song appearing among love songs). A single example stands out here nonetheless, the song *Si m destreignetz, dompna, vos et Amors* by Arnaut de Maroill falls outside the main author cluster, but is topically still consistent with the author's other works. Interestingly, this is one of the sole examples in this subset of the corpus where the attribution is uncertain, with the poem also ascribed to the otherwise unknown troubadour Ramont.

Some cluster overlaps are observed and can be interpreted in historical and linguistic contexts. The overlap between Uc de Saint-Circ and Bertran de Born, for instance, is likely influenced by their shared Limousin dialect, while the proximity of Raimbaut d'Aurenga and Marcabru, both early 12th-century troubadours, may reflect contemporaneous use of archaic Old Occitan. These patterns suggest that, beyond stylistic differences, genre and historical-linguistic factors may have a distinct effect on the embedding space.

Overall, the qualitative analysis supports the conclusion that the clustering predominantly reflects authorial style, with minor deviations attributable to genre differences or shared linguistic features.

4.1.3. Ablation study: The influence of Genre

Our analysis shows that poems which were erroneously assigned to the wrong author cluster often differ topically from the author's main oeuvre. This raises the question of whether genre confounds embedding-based clustering. Since transformer embeddings primarily capture semantic content, it is unclear if clusters reflect stylistic differences or mainly segregate works by genre. To test this, we partition the corpus into three coarse-grained genres: Love songs (n = 251), Satires (n = 96), and Religious songs (n = 33), and apply the same methodology using the best-performing monolingual and multilingual encoders. If genre drives the original clusters, performance on genre-restricted subsets should decrease relative to the scores in Table 2.

The results of these clustering experiments can be found in Table 3. We observe that when clustering within genre subsets, intrinsic metrics generally decline slightly, reflecting reduced variability in the smaller, more homogeneous subsets of data. Extrinsic metrics, however, improve notably for Love songs and Satires, indicating that embeddings capture stylistic differences when topical content is relatively controlled.

[3] https://github.com/JustGlowing/minisom

			Intrinsic			Extrinsic		
			SC ↑	CHI ↑	DB ↓	RI ↑	ARI ↑	MR ↓
Monolingual	camembert-base	**Love**	0.10	**40.6**	2.10	0.88	0.49	0.24
	camembert-base	**Satire**	0.12	22.57	2.05	0.83	0.45	**0.19**
	camembert-base	**Religion**	**0.19**	13.34	**1.48**	0.75	0.25	0.27
Multilingual	bert-base-multilingual-cased	**Love**	0.11	29.34	2.31	**0.89**	**0.57**	0.26
	bert-base-multilingual-cased	**Satire**	0.10	16.9	2.25	0.84	0.52	0.25
	bert-base-multilingual-cased	**Religion**	0.17	10.5	1.33	0.68	0.14	0.36

Table 3: Ablation study across three different topics: Love songs, Satires and Religious songs. For each genre we report the best-performing monolingual and multilingual encoder and the same evaluation metrics as in Table 2

Religious songs are an exception as both the intrinsic and extrinsic scores remain low. This genre exhibits high topical diversity – including devotion, blended love-religion themes, and Crusader songs – combined with conventional structures and fewer stanzas per author. These factors limit the distinctive lexical semantic features available to embeddings, making author separation significantly more difficult.

4.2. Document-level Authorship Attribution

While most studies in medieval authorship attribution focus on stylistic differences across longer documents, troubadour collaborations provide a chance to examine how distinct authorial styles manifest within a single poem. This allows us to explore fine-grained variation in style at the stanza level, rather than across entire works. From the 168 collaborative poems in the corpus, we select 89 for detailed analysis, filtering for sufficient length, at least two stanzas per author, and inclusion of historical authors rather than fictional interlocutors.

4.2.1. Methodology

To evaluate stylistic coherence in collaborative poems, we employ a pairwise similarity framework at the stanza level. For each poem, we compute similarity scores between all stanza pairs using a variety of traditional and modern metrics for modeling authorial style: (i) character n-grams, (ii) average line length, (iii) normalized line length, (iv) token-set Jaccard similarity, and (v) mBERT-based embeddings. Each metric yields a symmetric stanza-by-stanza similarity matrix $S \in R^{n \times n}$, where n denotes the number of stanzas in the poem.

Let $I_a \subseteq \{1, \ldots, n\}$ represent the set of stanza indices authored by writer a. Based on this partition, we compute two mean similarity scores for each author:

Intra-author similarity. The average similarity among stanza pairs written by the same author:

$$\text{Intra}(a) = \frac{1}{|P_{aa}|} \sum_{(i,j) \in P_{aa}} S_{ij},$$

$$P_{aa} = \{(i,j) \mid i,j \in I_a,\, i < j\}.$$

Cross-author similarity. The average similarity between stanzas written by author a and those written by the collaborating author b:

$$\text{Cross}(a) = \frac{1}{|P_{ab}|} \sum_{(i,j) \in P_{ab}} S_{ij},$$

$$P_{ab} = \{(i,j) \mid i \in I_a,\, j \notin I_a\}.$$

These scores provide a basis for quantifying stylistic distinctiveness within and across collaborators. In subsequent analyses, we collapse the two measures into a single comparative statistic, $\Delta(a) = \text{Cross}(a) - \text{Intra}(a)$, which captures whether an author's stanzas are closer in style to their own writing or to that of their collaborator(s).

4.2.2. Results and Analysis

Quantitative Results Building on the intra- and cross-author similarity measures introduced above, we aggregate these into a single comparative statistic, Δ, which can be defined as the difference between cross- and intra-author similarity:

$$\Delta(a) = \text{Cross}(a) - \text{Intra}(a).$$

This statistic indicates whether stanzas by an author are more similar to their own writing ($\Delta < 0$) or to their collaborator ($\Delta > 0$). At the corpus level, we aggregate Δ by averaging across all authors and poems, yielding a single value per metric that summarizes its ability to distinguish intra- from cross-author similarity.

Table 4 reports the average Δ across collaborative poems. Character n-grams show the strongest, most consistent author-specific signal,

while embedding-based cosine similarity is effective but more variable. Token-based metrics (type-token ratio, Jaccard similarity) show moderate sensitivity, and average line length performs poorly. Overall, simple distributional features and especially character n-grams seem to outperform embeddings in unsupervised, fine-grained settings.

Metric	Average Δ
Character n-grams	-0.0222
Character skip-grams	-0.02
Average line length	0.01
Jaccard similarity	-0.013
Embedding similarity	-0.003

Table 4: Average Δ across all collaborations in the corpus. Negative values indicate greater similarity to the author's own writing, while positive values indicate greater similarity to the collaborator.

Qualitative Results In addition to quantitative comparisons of mean similarity differences, we also conduct a qualitative analysis at the level of individual collaborations. For each metric and collaboration, we examine whether authors' stanzas are closer to their own writing than to their collaborator's. Specifically, we use the difference statistic $\Delta(a)$ defined in the previous section.

Stanza-level similarity can be characterized in three ways: **full alignment**, where stanzas by both authors exhibit higher similarity to their own works; **partial alignment**, where only one author's stanzas are more similar to their own, while the other aligns more closely with another's; and **cross alignment**, where stanzas by both authors display greater similarity to those of the other author. For a given collaboration involving two authors a_1 and a_2, we then formally define the outcome function:

$$(a_1, a_2) = \begin{cases} \text{Full} & \Delta(a_1) < 0 \wedge \Delta(a_2) < 0, \\ \text{Partial} & \Delta(a_1) < 0 \oplus \Delta(a_2) < 0, \\ \text{Cross} & \Delta(a_1) > 0 \wedge \Delta(a_2) > 0. \end{cases}$$

We then count the outcomes across all collaborations and all metrics. The counts provide an interpretable diagnostic: they indicate whether authors' stanzas exhibit consistent stylistic self-similarity within collaborations, or whether their style is less distinguishable from that of their collaborator. To improve the interpretability of the results, we conduct statistical significance testing on the distribution of count labels in order to evaluate the effectiveness of the proposed metrics for modeling authorial style. We test the N_{null} that no systematic stylistic differences exist between the authors. To approximate the null distribution, we employ a permutation randomization test (n = 1000), in which stanza–author labels are randomly reassigned within each poem. This provides a baseline against which the observed results can be compared, allowing us to assess whether the detected similarities or alignments are unlikely to have arisen by chance.

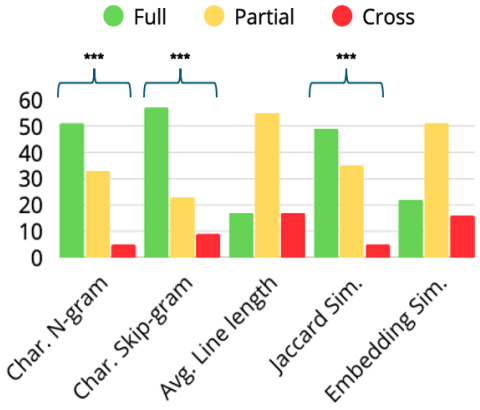

Figure 4: Counts of stylistic consistency inside of collaborative poems. Significant results deviating from the expected counts are highlighted

Based on the expected counts under N_{null}, we perform a chi-square test for each of the five evaluated metrics. The outcomes of these tests are visualized in Figure 4. Significant results are obtained for character-level n-grams ($p = 2.8^{-18}$), character-level skip-grams ($p = 2.2^{-22}$,), and the Jaccard Index ($p = 3.6^{-15}$), indicating a clear distinction between authors within the same poem. These findings suggest that the three character-level metrics are the most effective for modeling stylistic variation in this context.

By contrast, line length ($p = 0.37$) and embedding-based similarity ($p = 0.12$) are not significant. Our results show that line length, traditionally robust in stylometry, is apparently less suitable for troubadour collaborations due to strict formal conventions, rhyme and syllabic meter, producing superficially similar stanzas. Similarly, embeddings fail at the stanza level because collaborative poems revolve around shared themes, aligning poets' semantic spaces, even though embeddings work at the corpus level.

5. Conclusion

This study introduced TrovaCor, the largest publicly available corpus of medieval Occitan troubadour poetry designed for computational analysis. Through document- and stanza-level authorship attribution experiments, we showed that both traditional distributional metrics and modern transformer-based embeddings can capture meaningful stylistic differences, even in a low-resource

historical setting. While character-level features remain strong baselines, multilingual embeddings enable cross-author comparisons that extend beyond topical cues. Our analysis further highlights how genre and formulaicity complicate stylistic clustering, underscoring the need for genre-aware approaches to historical stylometry. By releasing TrovaCor and providing a transparent methodological framework, we aim to support future research in authorship attribution, genre modeling, and the broader computational study of medieval traditions.

6. Bibliographical References

2025. Gallica: Bibliothèque numérique de la bibliothèque nationale de france. https://gallica.bnf.fr. Digital library of the BnF and its partners.

2025. Project gutenberg. https://www.gutenberg.org. Online digital library / e-book repository.

Carl Ludwig Ernst Appel. 1892. *Provenzalische inedita aus Pariser handschriften*, volume 13. OR Reisland.

Esteban Garces Arias, Vallari Pai, Matthias Schöffel, Christian Heumann, and Matthias Aßenmacher. 2023. Automatic transcription of handwritten old occitan language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15416–15439.

Stefano Asperti and Luca De Nigro. 2003. Bedt 02-15. Online. Accessed: 2025-10-01.

Karl Bartsch. 1872. *Grundriss zur Geschichte der provenzalischen Literatur*. RL Friderichs.

Pierre Bec. 1982. Le problème des genres chez les premiers troubadours. *Cahiers de civilisation médiévale*, 25(97):31–47.

Pierre Bec. 1984. Burlesque et obscénité chez les troubadours: pour une approche du contre-texte médiéval. *(No Title)*.

Jasmin Bharadiya. 2023. A comprehensive survey of deep learning techniques natural language processing. *European Journal of Technology*, 7(1):58–66.

Christelle Chaillou. 2009. La poésie lyrique des troubadours. musique, poésie, contexte. *Annales de Vendée*, (4):139–157.

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

Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. 2019. The epistle to cangrande through the lens of computational authorship verification. In *New Trends in Image Analysis and Processing–ICIAP 2019: ICIAP International Workshops, BioFor, PatReCH, e-BADLE, DeepRetail, and Industrial Session, Trento, Italy, September 9–10, 2019, Revised Selected Papers 20*, pages 148–158. Springer.

Jeroen De Gussem, Samu Niskanen, and James Willoughby. 2022. Computational stylistics and medieval texts. In *Routledge resources online: medieval studies*, pages 1–12. Routledge.

Loic De Langhe, Orphée De Clercq, and Veronique Hoste. 2024. Unsupervised authorship attribution for medieval latin using transformer-based embeddings. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 57–64.

Nesibe Merve Demir. 2013. Artificial neural network techniques in authorship attribution. *Southeast Europe Journal of Soft Computing*, 2(2).

Institut d'Estudis Catalans. 2020. Corpus des troubadours. Digital Corpus of Old Occitan. Accessed: 2025-10-01.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Costanzo Di Girolamo et al. 2012. *Rialto. Repertorio informatizzato dell'antica letteratura trobadorica e occitana, a cura di C. Di Girolamo. http://www. rialto. unina. it. 2003 e segg. Sito web. Pubblicazione periodica permanente iniziata nel 2003. ISSN 1973-381X*. Università di Napoli Federico II. Dipartimento di Filologia Moderna.

Gui d'Ussel. 1922. *Les poésies des quatre troubadours d'Ussel: publiées d'après les manuscrits*. Delagrave.

Thomas T Field. 2006. Troubadour performance and the origins of the occitan" koine". *Tenso*, 21(1):36–54.

Vanessa B Gorman and Robert J Gorman. 2016. Approaching questions of text reuse in ancient greek using computational syntactic stylometry. *Open Linguistics*, 2(1).

David I Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, 13(3):111–117.

Fereshteh Jafariakinabad. 2021. Machine learning techniques for topic detection and authorship attribution in textual data.

Jurgita Kapočiūtė-Dzikienė, Andrius Utka, and Ligita Šarkutė. 2015. Authorship attribution of internet comments with thousand candidate authors. In *Information and Software Technologies: 21st International Conference, ICIST 2015, Druskininkai, Lithuania, October 15-16, 2015, Proceedings 21*, pages 433–448. Springer.

Andrew Kehler and Andreas Stolcke. 1999. Unsupervised learning in natural language processing. In *Association for Computational Linguistics. Proceedings of the workshop. In Preface A. Kehler and A. Stolcke, editors*.

Mike Kestemont. 2012. Stylometry for medieval authorship studies: an application to rhyme words. *Digital Philology: A Journal of Medieval Cultures*, 1(1):42–72.

Mike Kestemont, Sara Moens, and Jeroen Deploige. 2015. 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.

Kathryn Klingebiel. 1997. Lost literature of the troubadours: A proposed catalogue. *Tenso*, 13(1):1–23.

Teuvo Kohonen. 2013. Essentials of the self-organizing map. *Neural networks*, 37:52–65.

Robert Layton, Paul Watters, and Richard Dazeley. 2013. Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering*, 19(1):95–120.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Fernanda López-Escobedo, Carlos-Francisco Méndez-Cruz, Gerardo Sierra, and Julián Solórzano-Soto. 2013. Analysis of stylometric variables in long and short texts. *Procedia-Social and Behavioral Sciences*, 95:604–611.

Carolina Martín-del Campo-Rodríguez, Grigori Sidorov, and Ildar Batyrshin. 2022. Unsupervised authorship attribution using feature selection and weighted cosine similarity. *Journal of Intelligent & Fuzzy Systems*, 42(5):4357–4367.

Roser Morante, Eleanor LT Smith, Lianne Wilhelmus, Alie Lassche, and Erika Kuijpers. 2022. Identifying copied fragments in a 18th century dutch chronicle. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5865–5878.

Frederick Mosteller and David L Wallace. 1963. 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.

Antonio Neme, JRG Pulido, Abril Muñoz, Sergio Hernández, and Teresa Dey. 2015. Stylistics analysis and authorship attribution algorithms based on self-organizing maps. *Neurocomputing*, 147:147–159.

Erkki Oja and Samuel Kaski. 1999. *Kohonen maps*. Elsevier.

William D. Paden. 1998. *An Introduction to Old Occitan*, volume 4 of *Introductions to Older Languages*. Modern Language Association of America, New York.

Polina Panicheva, Olga Litvinova, and Tatiana Litvinova. 2019. Author clustering with and without topical features. In *International Conference on Speech and Computer*, pages 348–358. Springer.

Tie-Yun Qian, Bing Liu, Qing Li, and Jianfeng Si. 2015. Review authorship attribution in a similarity space. *Journal of Computer Science and Technology*, 30(1):200–213.

RVSPK Ranatunga, AS Atukorale, and KP Hewagamage. 2011. Intrinsic plagiarism detection with kohonen self organizing maps. In *U The International Conference on Advances in ICT for Emerging Regions-ICTer2011*, volume 125.

Patrick Sauzet. 2012. Occitan plurals: A case for a morpheme-based morphology. In *Inflection and word formation in Romance languages*, pages 179–200. John Benjamins Publishing Company.

Jacques Savoy. 2020. Machine learning methods for stylometry. *Cham: Springer*.

Matthias Schöffel, Marinus Wiedner, Esteban Garces Arias, Paula Ruppert, Christian Heumann, and Matthias Aßenmacher. 2025. Modern models, medieval texts: A pos tagging study of old occitan. *arXiv preprint arXiv:2503.07827*.

Olga Scrivner and Sandra Kübler. 2012. Building an old occitan corpus via cross-language transfer. In *KONVENS*, pages 392–400.

Lisa Shields. 1976. Medieval manuscripts in french and provençal. *Hermathena*, pages 90–100.

Jean Sibille. 2024. Les dialectes occitans. *Manuel de linguistique occitane*, pages 423–471.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Wolf-Dieter Stempel, Helmut Stimm, Claudia Kraus, Renate Peter, and Monika Tausend. 1996. *Dictionnaire de l'occitan médiéval: DOM*. Niemeyer.

Justin Anthony Stover, Yaron Winter, Moshe Koppel, and Mike Kestemont. 2016. Computational authorship verification method attributes a new work to a major 2nd century a frican author. *Journal of the Association for Information Science and Technology*, 67(1):239–242.

Ludovic Tanguy, Franck Sajous, Basilio Calderone, and Nabil Hathout. 2012. Authorship attribution: Using rich linguistic features when training data is scarce. In *PAN Lab at CLEF*.

Fiona J Tweedie, Sameer Singh, and David I Holmes. 1996. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, 30:1–10.

Caroline Vandyck, Wouter Haverals, and Mike Kestemont. 2025. Making characters count. a computational approach to scribal profiling in 14th-century middle dutch manuscripts from the carthusian monastery of herne. *arXiv preprint arXiv:2509.00067*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Christin Michelle Laroche Wilson. 2012. *Variation and Text Type in Old Occitan Texts*. The Ohio State University.

Sam Wolfe. 2018. Occitan, verb second and the medieval romance word order debate. In *Romance languages and linguistic theory 13: Selected papers from 'Going Romance'29, Nijmegen*, pages 315–336. John Benjamins Publishing Company.

Appendix

A. Detailed Genre Distribution

Coarse-Grained Genre	Fine-Grained Genre	Number of Documents	Number of Tokens
Love Songs	Canso	768	253,675
	Pastorella	28	10,445
	Salut d'Amor	11	9,108
	Romance	5	2,582
	Alba	5	1,016
	Sonnet	3	324
	Serena	1	183
Satires	Sirventes	319	97,405
	Canso-sirventes	4	1,177
	Satira	2	748
	Miei-sirventes	2	441
	Gap	1	393
	Trufa	1	367
Collaborations	Tenso	68	19,967
	Tenso (Partimen)	49	16,942
	Exchange of Coblas	35	5,049
	Fictional Tenso	16	5,286
Religious Songs	Chanson Religieuse	52	16,454
	Chanson de Croisade	30	11,679
	Alba Relegieuse	6	2,134
	Preghiera	2	1,738
Short Verse	Cobla	87	5,199
	Coblas	24	2,978
	Coblas with Tornada	17	2,560
	Cobla with Tornada	5	606
Mourning Songs	Plahn	35	14,318
Dancing Songs	Dansa	17	2,268
	Estampida	5	1,788
	Balada	3	527
Other	Vers	18	5,783
	Descort	17	5,257
	Retroencha	3	965
	Plazer	2	449
	Escondich	2	634
	Sestina	2	568
	Devinaill	2	558
	Ensenhamen	2	4,223
	Epistola	1	4,513

Table 5: Dataset Statistics by Genre

B. Genre Gloss

Canso	Classic courtly love song of the troubadours, expressing refined love (fin'amor) for a usually unattainable lady; formal stanzaic structure.
Alba religieuse	Devotional adaptation of the alba (dawn song), using the dawn motif for spiritual or moral themes.
Cobla	Single stanza of a poem; can also be an independent short composition.
Coblas	Plural of cobla; stanzas of a poem, often with variation in rhyme schemes.
Tenso	Poetic debate between two troubadours on moral, political, or amorous topics.
Sirventes	Moral, political, or satirical song modeled on the canso but addressing ethical or social issues.
Chanson religieuse	Pious or devotional song, often praising the Virgin Mary or Christian virtue.
Planh	Lament or elegy mourning the death of an important person, patron, or loved one.
Estampida	Lively rhythmic dance song or instrumental form; dance-oriented lyric composition.
Salut d'amor	Long didactic or rhetorical "love greeting" poem in epistolary form.
Dansa	Dance song with a refrain, light in tone for performance with movement.
Vers	General term for a lyric poem, often moral, didactic, or philosophical.
Tenso (Partimen)	Structured tenso where one poet poses a dilemma and the other chooses and defends a side.
Pastorela	Narrative lyric about a knight meeting a shepherdess, typically involving flirtation or debate.
Cobla avec tornada	Stanza including a short final refrain addressed to a person or audience.
Chanson de croisade	Crusade song expressing support or reflection on Crusades; themes of faith and separation.
Échange de coblas	Poetic exchange of stanzas between poets, often in dialogue form.
Alba	Dawn song where lovers lament separation at daybreak, often featuring a watchman's warning.
Descort	"Discordant" poem with irregular stanzas, rhyme, or meter, reflecting emotional or thematic disunity.
Tenson fictive	Fictional tenso, where a single poet writes both sides of a debate.
Carta, Epistola en vers	Letter written in verse, often moral, didactic, or amorous.
Plazer	Poem celebrating joy or pleasure, sometimes counterpoint to the planh.
Balada	Dance song with a refrain, lighter or more popular in tone.
Escondich	Poem where the lover conceals or denies his love.
Serena	Evening song, opposite of alba; lover anticipates nightfall.
Miei-sirventes	Hybrid form combining sirventes and canso elements.
Sonnet	14-line poem; later import from Italian models.

Sestina	Highly formal poem of six 6-line stanzas plus envoi, repeating six end-words; invented by Arnaut Daniel.
Devinaill	Riddle poem or enigmatic composition, often playful or allegorical.
Preghiera	Prayer in verse; devotional lyric addressed to God or a saint.
Romance	Narrative lyric or short tale in verse, often recounting love adventures or legends.
Gap	Boast or challenge poem asserting superiority, inviting rivals to respond.
Retroencha	Poetic response or continuation of another poet's work.
Satira	Satirical poem targeting social, moral, or personal faults, akin to sirventes.
Sirventes-Canso	Mixed form combining sirventes (moral/political) and canso (love).
Ensenhamen	Didactic poem offering moral, ethical, or practical instruction.
Trufa	Mocking or humorous poem, often parodying courtly or religious conventions.