Revisiting Authorship Attribution of *Tirant lo Blanc* Using Parts of Speech *n*-grams

Yoshifumi Kawasaki
University of Tokyo
ykawasaki@g.ecc.u-tokyo.ac.jp

Abstract

Tirant lo Blanc (TLB) is a masterpiece of medieval Catalan chivalric romance. Regarding its authorship, two hypotheses exist: the single-authorship hypothesis claims in agreement with the dedication that Joanot Martorell is the sole author, whereas the dual-authorship hypothesis alleges in line with the colophon that Martorell wrote the first three parts and Martí Joan de Galba added the fourth part. In this study, we revisit the unsettled authorship attribution of TLB with stylometric techniques; specifically, we exploit parts-of-speech (POS) *n*-grams as stylistic features to investigate stylistic differences (if any) across the work. Furthermore, we address the distinction between narration and conversation, which has previously been omitted. We performed exploratory multivariate analyses and demonstrated that, despite internal differences, single-authorship is more likely from a statistical point of view. If Galba had contributed something to the last quarter of the work, it would have been minimal.

1 Introduction

Tirant lo Blanc, hereafter TLB, is a chivalry novel written in Catalan toward the end of the 15th century. Its first edition was printed in 1490 in Valencia, Spain, although it had been supposedly composed between 1460 and 1465 (Ferrando, 2012). The Medieval Catalan literary masterpiece was praised for its style as "the best book in the world" by the 17th-century Spanish writer Cervantes in chapter VI of *Don Quijote de la Mancha* (de Cervantes Saavedra, 1999). The work was deemed to be the very first modern novel in Europe by Peruvian Nobel laureate Mario Vargas Llosa, who rediscovered and acknowledged its literary values in recent times (Vargas Llosa, 2015).

Regarding its authorship, a sharp contradiction exists between the dedication at the beginning of the book and the colophon at its end, where information about the authorship and printing is provided. Joanot Martorell affirms in the dedication that he is solely responsible for the work (single-authorship hypothesis), whereas the colophon states that Martorell *translated* the first three of the four parts and that the fourth part was *translated* by Martí Joan de Galba (dual-authorship hypothesis). Here, *translation* refers to creation, as feigning a *translation* was commonplace during the period under consideration. The apparent inconsistency has been reconciled supposing that Martorell wrote most of the work and Galba revised and expanded it later (Martorell, 2008). Nonetheless, this supposition requires empirical validation to verify whether Galba actually participated in the creation and, if so, to identify Galba's contributions.

Thus, this study revisits the unsettled authorship attribution of TLB by exploiting parts-of-speech (POS) *n*-grams as stylistic features. Existing literature has only considered a word-length distribution, that of the most frequently used words, and indices of the diversity of vocabulary. This study also addresses the distinction between narration and conversation, which has hitherto been disregarded, for fear that varying proportions of the two components in the work might confound the eventual outcome. We performed exploratory multivariate analyses and demonstrated that, despite internal differences, single authorship is more likely from a statistical point of view. If Galba had contributed something to the last quarter of the work, it would have been minimal.

The remainder of this paper is organized as follows. In Section 2, we review the existing literature and highlight its limitations. Section 3 describes the methodology followed in this study. In Section 4, we present the experimental results, followed by a discussion in Section 5. Finally, Section 6 concludes the paper.

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 16–26

2 Related Work

The single-authorship hypothesis, according to which Martorell is the sole author, is based on the description in the dedication, whereas the dual-authorship hypothesis, according to which Martorell wrote the first three parts and Galba added the fourth, is indicated in the colophon. The single-authorship hypothesis has been endorsed by renowned philologist Martí de Riquer (de Riquer, 1990; Martorell, 2016), although the dual-authorship hypothesis has not been completely rejected (Martorell, 2008). Assuming that the dual-authorship hypothesis is true, the question arises as to where the fourth part that Galba purportedly composed begins. TLB is not explicitly divided into four parts, except for the first part, the beginning of which is noted ahead of chapter 1. Considering that TLB consists of 487 chapters of unequal length, de Riquer (1990) estimated that if the colophon is to be trusted, the fourth part begins with chapter 363 in terms of the number of chapters, or around chapter 283 in terms of the total length of the novel.

Under these circumstances, stylometry plays a key role (Martorell, 2008). Stylometry aims to identify the genuine author(s) of a written text through quantitative analysis (Stamatatos, 2009). A series of stylometric studies have delved into questions concerning the authorship of TLB. In their pioneering study, Girón et al. (2005) examined the distribution of word lengths (Mendenhall, 1887; Williams, 1975) and the most frequent context-free words, including articles, conjunctions, prepositions, and pronouns. They detected a change in the distribution of the variables from chapters 371 to 382 and concluded that the results corroborate dual authorship. Nonetheless, they admit the possibility that the observed differences may have been due to factors other than changes in authors.

One shortcoming of Girón et al. (2005) is that word-length distribution is not currently viewed as the most effective feature for authorship attribution tasks, which is implied by its practical absence in recent studies. Furthermore, the linguistic interpretation of word-length distribution is not straightforward. The stylistic information encoded therein is unclear.

Another drawback is the model selection process for the number of authors involved in the work. They compared two probabilistic models corresponding to the single and dual authorship hypotheses. The former consists of a single multinomial distribution, and the latter comprises a mixture of two distributions. Then, the ratio of posterior probabilities between the two models was computed to decide which one was more likely. However, they did not consider the model's complexity. Hence, the selection of the dual authorship hypothesis was the natural outcome, given that a more complex model fits better than a simpler one. The trade-off between model complexity and goodness of fit should be addressed appropriately.

In addition, the authors disregarded the distinction between narration and conversation when the analyses were vulnerable to the varying proportions of these two components in the work. In fact, the narration/conversation ratio fluctuates greatly among chapters, as depicted in Figure 1. The vertical axis represents the narration ratio, computed as the number of tokens in the narration divided by the chapter length. The curve represents the moving average with a window size of 20. The ratio of narration remains high from around chapter 375 onward to the end, whereas it is negligible from chapters 40–100. We assume that a different proportion of narration/conversation is not *per se* indicative of different authorship because its constancy across a work by a single author is not self-evident; narration/conversation may well be abundant in some sections and exiguous in others.

Using analogous approaches, other studies arrived at the same conclusion (Girón et al., 2005; Riba and Ginebra, 2005; Puig et al., 2015; Font et al., 2016). Riba and Ginebra (2006) also reinforces their findings using eight different indices of the diversity of vocabulary, which is rarely utilized as an effective stylistic feature either.

Thus, this study intends to shed new light on the authorship attribution of TLB in the following ways: (i) we leverage POS *n*-grams, which are effective and linguistically interpretable stylistic features; (ii) we conduct model selection appropriately, considering the trade-off between model complexity and goodness of fit; and (iii) we address the distinction between narration and conversation, which has hitherto been omitted.

3 Methods

A digitized transcription of the *princeps* edition was used in this study (Martorell, 2006)[1]. The

[1] https://www.cervantesvirtual.com/obra/tirant-lo-blanc--1/

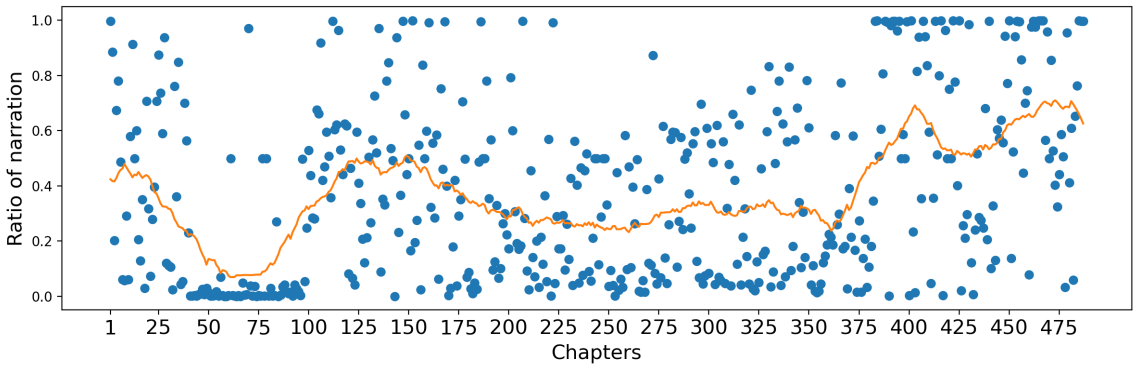

Figure 1: Ratio of narration along the chapters. The ratio was computed as the number of tokens in the narration divided by the chapter length. The curve represents the moving average with a window size of 20.

chapter titles and Latin phrases (e.g., *deo gracias* "thanks to God") were removed. We also eliminated paragraphs in the letter format that deviated from the typical structure of the work. Numerous passages allegedly plagiarized from other works (de Riquer, 1990) were retained as such for the sake of simplicity. Regarding punctuation, commas were eliminated so that editorial interventions would not come into play, whereas periods, colons, semi-colons, interrogation marks, and exclamation marks indicating sentence boundaries were retained as single punctuation symbols. Moreover, contracted and concatenated forms were separated prior to POS tagging (e.g., *l'art* "the art" and *donant-lo* "giving it" were divided into *l' art* and *donant -lo*, respectively).

Pre-processing resulted in 420,879 tokens and a vocabulary size of 17,181. For subsequent analyses, we did not adopt the original chapter division because the lengths varied considerably from one another. Instead, we generated equal-length pieces of 10K tokens to obtain reliable statistics. The length of 10K tokens is way above the minimum sample size of 5K tokens that was shown to be sufficient for stylometric analysis (Eder, 2015). The shortest final piece of 879 tokens was merged into the penultimate one. Thus, the entire work resulted in forty-two pieces.

We leveraged POS *n*-grams as stylistic features. The effectiveness of POS *n*-grams has been demonstrated by the previous research addressing literary works in multiple languages, including English (Koppel et al., 2002; Clement and Sharp, 2003; Juola, 2006; Hirst and Feiguina, 2007; Eder, 2015; Pokou et al., 2016; Savoy, 2017), French (Kocher and Savoy, 2019), Japanese (Uesaka and Murakami, 2015), and Spanish (Kawasaki, 2021, 2022). The advantages of employing POS sequences are multi-fold: (i) the numerous occurrences provide reliable statistics; (ii) they are relatively, if not completely, independent of content; (iii) they are deemed to be reliable style markers (Holmes, 1998; Juola, 2006; Stamatatos, 2009). Although partially, they capture syntactic patterns that are difficult to imitate and allegedly out of the conscious control of the author (Baayen et al., 1996); and (iv) they are supposedly less vulnerable to editorial interventions that would manipulate the original; in fact, orthographic vacillation could derive not only from the author but also from the typesetters in the Middle Ages.

The tokens were POS-tagged according to lemmatized concordance[2]. Specifically, we looked up each token in the concordance prepared in keywords in context format, considering its preceding and following contexts. Thus, more than 99% of the tokens were correctly tagged. Tokens that were ambiguous or left untagged were assigned a special tag, UNK, for simplicity, although manual tagging was desirable. Consequently, the number of POS tags amounted to thirteen[3]. For the most frequent twenty words, including adverbs, conjunctions, prepositions, and verbs, we adopted lemma forms in lieu of the POS-tags to exploit their particular usage[4]. For example, the preposition *de* "of"

[2]We are greatly indebted to Dr. Eduard Baile López of University of Alicante for providing us with the valuable data.

[3]ADJ(ECTIVE), ADV(ERB), ART(ICLE), CONJ(UNCTION), CONTR(ACTION BETWEEN PREPOSITION AND ARTICLE), INTERJ(ECTION), N(OUN), PREP(OSITION), PRON(OUN), PROPER(NOUN), PUNCT(UATION), UNK(NOWN), and V(ERB)

[4]*i* "and", *de* "of", *que* "that", *ésser* "to be", *en* "in", *a* "to", *per* "for", *no* "not", *fer* "to do", *haver* "to have", *tot* "all", *com* "as", *ab* "with", *dir* "to say", *molt* "much", *se* "oneself", *gran* "great", *un* "a", *qui* "who", and *tenir* "to have".

was not converted into PREP but maintained as such. This resulted in thirty-three unigram types in total: thirteen POS tags and twenty lemma forms.

For the subsequent multivariate analyses, every text piece was represented as a vector, with its elements being the z-transformed relative frequencies of the n-grams (Burrows, 2002). The relative frequencies were standardized to have a zero mean and unit variance for every variable. We considered only the most frequent POS n-grams above a given rank threshold r, whereas the remainder was aggregated under the OTHERS label. Thereafter, we performed two exploratory multivariate analyses, i.e., principal component analysis (PCA) and k-means clustering. As no other works by the relevant authors were available, it was infeasible to apply supervised methods such as classification. To assess the robustness of the analyses, we varied the n-gram size n for $n \in \{1, 2, 3, 4\}$ and the rank threshold r for $r \in \{50, 100, 300, 500\}$. For $n = 1$, r was fixed at 33, which is the number of unigram types.

4 Results and Analyses

In this section, we present the experimental results without a narration/conversation distinction. The results of the respective parts will be presented in Section 5.1. For illustrative purposes, we provide the results obtained with the hyper-parameters $(n, r) = (3, 300)$, unless noted otherwise.

First, we examined the overall similarity patterns across the entire work. Figure 2 displays the pair-wise distance scores between the pieces. The i-th piece is designated as TLB_i. The scores were calculated as $\sqrt{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2 / r}$, where $\boldsymbol{x_i}$ represents the feature vector for the i-th piece and r the number of n-grams considered. The bluer (redder) the cell, the more (less) similar the pair of pieces. We can readily discern a large cluster comprised of TLB_01–TLB_34, within which the distance scores are small compared to the rest of the pieces.

4.1 PCA

We performed PCA using `sklearn.decomposition.PCA` with the default settings (Pedregosa et al., 2011)[5]. Figure 3 illustrates the first two PC scores obtained with the hyper-parameters $(n, r) = (3, 300)$. The

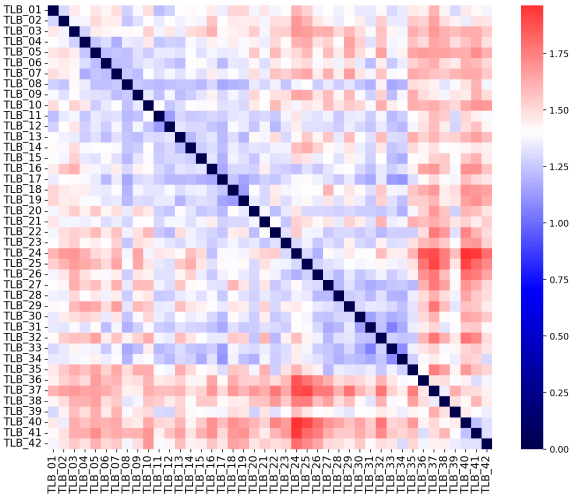

Figure 2: Pair-wise distance scores between the 10K-token pieces from the entire work, computed with hyper-parameters $(n, r) = (3, 300)$. The bluer (redder) the cell, the more (less) similar the pair of pieces.

contribution ratios of PC1 and PC2 were 12.3% and 9.8%, respectively. Figure 3 apparently shows no significant pattern. However, we found that both PC1 and PC2 presented a moderate negative correlation with the proportion of conversational parts in the pieces: Spearman's $\rho = -0.66 \quad (p < 0.01)$ for PC1 and $\rho = -0.34 \quad (p = 0.03)$ for PC2. It is probable that the principal components simply reflect the proportion of narration/conversation, although it is not impossible that they reflect different authorship. Hence, we find it more practical to distinguish between the two parts and verify whether the same pattern emerges.

4.2 *k*-means

We performed k-means clustering using `sklearn.cluster.KMeans` with the default settings (Pedregosa et al., 2011)[6]. The number of clusters k was fixed at $k = 2$, which is the supposed maximum number of authors involved. As the algorithm is sensitive to the initially selected centroids, we ran it 100 times to compute the mean concordance rate, which is defined as the average number of times a pair of pieces is found in the same cluster. Our premise was that no clear-cut pattern would emerge if stylistic differences did not exist.

Figure 4 presents the pair-wise mean concordance rates obtained with hyper-

[5] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

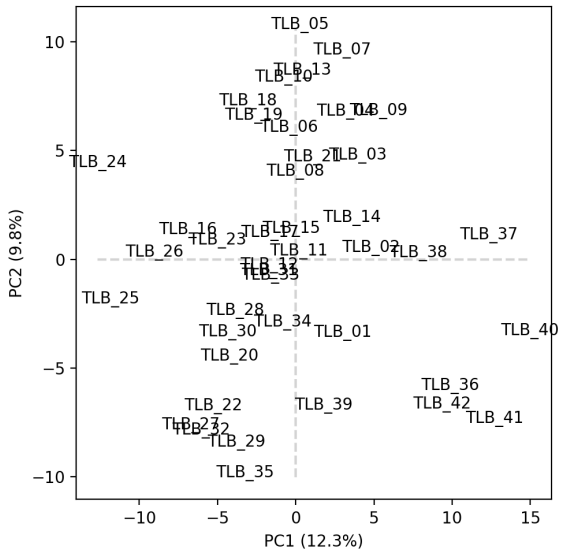

Figure 3: Scatter plot of PC1 and PC2 for the entire work with hyper-parameters $(n, r) = (3, 300)$.

parameters $(k, n) = (2, 3)$, while varying $r \in \{50, 100, 300, 500\}$. The darker the cell, the more often the pair of pieces were classified into the same cluster. Figure 4 illustrates that the clustering method is susceptible to the hyper-parameter r, resulting in inconsistent outcomes. The resulting clusters were also sensitive to n (data not shown). Consequently, it was difficult to draw definitive conclusions. If two distinct styles were to exist in the work, they would have been detected consistently regardless of different hyper-parameter settings, which was not the case.

For $r \in \{300, 500\}$, we can see a boundary between TLB_35 and TLB_36, which agrees with the findings of Riba and Ginebra (2005). They detected it in chapters 371–382, which roughly correspond to the second half of TLB_35 and the first half of TLB_36. However, this is also the point where the narration ratio increases (Figure 1). Therefore, we suspect that what Riba and Ginebra (2005) detected was not necessarily a change-point of authors but rather that of the narration/conversation ratio, and argue for the distinction between narration and conversation parts.

5 Discussion

5.1 Narration/Conversation Discrimination

As described above, the unequal amount of narration/conversation in the work potentially affects the resultant n-gram distribution. To avoid possible confounding effects, we distinguished between the narration and conversation sections. Identification of the two parts was readily made as the beginning of the conversation paragraphs is indicated with special characters. The entire text was first segregated into narration and conversation parts, and then each part was divided into 10K-token pieces. When the length of the last piece exceeded 6K, it was treated as an independent piece; otherwise, it was merged into the penultimate piece to prevent it from suffering data paucity. Thus, the narration and conversation parts resulted in eighteen and twenty-four pieces, respectively. The i-th piece in the narration (conversation) part was designated as TLB_N(C)_i.

5.1.1 Narration

Figure 5 illustrates the first two PC scores for the narration part with the hyper-parameters $(n, r) = (3, 300)$. The contribution ratios of PC1 and PC2 were 18.0% and 10.2%, respectively. PC1 neatly separates TLB_N_14–TLB_N_18 on the far left side from the rest, whereas the interpretation of PC2 is difficult to make.

The pair-wise mean concordance rates among the narration parts are displayed in Figure 6. The results were relatively robust with other hyper-parameter settings. The narration section presents a clear boundary between TLB_N_13 and TLB_N_14, which approximately corresponds to chapter 350, where the story turns to the fate of *Plaerdemavida*. This boundary does not diverge greatly from the estimation by de Riquer (1990) that the beginning of the fourth part should be situated in chapter 363 in terms of the number of chapters. Furthermore, it accords with de Riquer's earlier opinion that Galba's contribution should be located from chapter 349 onward (Martorell and de Galba, 1947). In sum, the detected boundary does not contradict the description in the colophon that Galba created the fourth section.

5.1.2 Conversation

Figure 7 illustrates the first two PC scores for the conversation part with the hyper-parameters $(n, r) = (3, 300)$. The contribution ratios of PC1 and PC2 were 18.2% and 10.6%, respectively. PC1 separates TLB_C_02–TLB_C_06 on the far left side from the rest, and among which PC2 isolates TLB_C_22–TLB_C_24 from the remainder.

Next, we present the pair-wise mean concordance rates for the conversation component in Figure 8. The conversation part presents two bound-

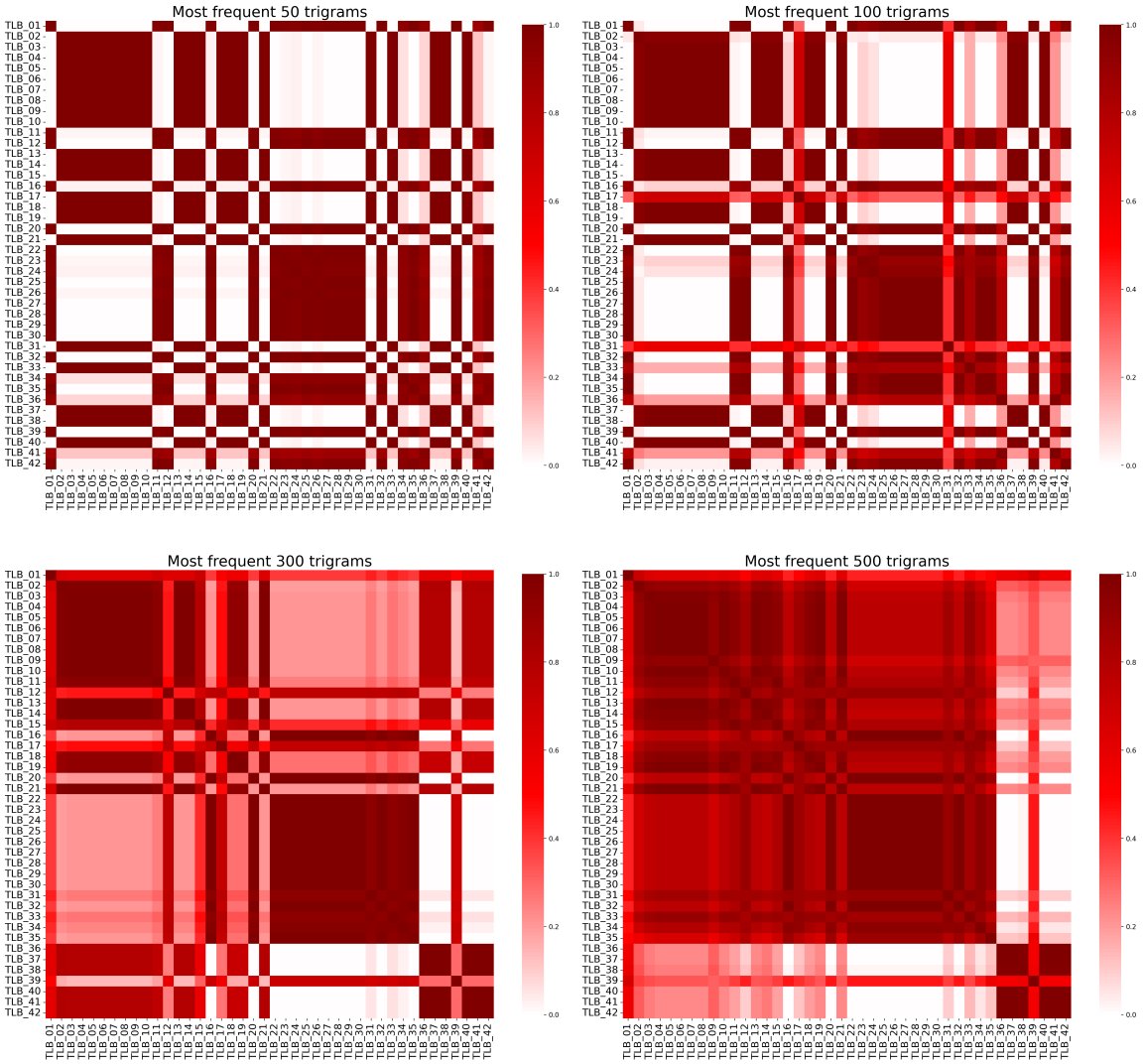

Figure 4: Pair-wise mean concordance rates computed from 100 iterations of k-means. The hyper-parameters were set at $(k, n) = (2, 3)$ and $r \in \{50, 100, 300, 500\}$. The darker the cell, the more similar the pair of pieces.

aries: one between TLB_C_01 and TLB_C_02, which corresponds approximately to chapter 29, and the other between TLB_C_06 and TLB_C_07, which corresponds approximately to chapter 101. The results were relatively robust with other hyper-parameter settings.

The pieces TLB_C_02–TLB_C_06, or chapters 29–101, roughly correspond to the latter part of the section "William of Warwick" and the entire section of "Tirant in England" (de Riquer, 1990). These chapters are exceptional in that they consist of conversation only (Figure 1) and are characterized by an abundance of narrational components within conversation, in contrast to the dialogic style of the rest of the chapters. This peculiarity could be attributed to the alleged adaptation for TLB of *Guillem de Vàroich* (GV), which Martorell himself would have composed prior to the creation of TLB (Gili i Gaya, 1947; de Riquer, 1990)[7]. In such a case, the second boundary between TLB_C_06 and TLB_C_07 would not necessarily reflect different authorship but rather Martorell's internal stylistic variation.

Regarding the first boundary between TLB_C_01 and TLB_C_02, it is noticeable that TLB_C_01 corresponding to the first part of "William of Warwick" does not resemble its continuation but the rest of the work starting from TLB_C_07. We speculate that Martorell's intensive retouching of the aforementioned GV only involved its initial part to accommodate it to the newly composed TLB and that the rest was left relatively intact.

Notably, the narration and conversation diverged in terms of the boundary that separates the two in-

[7] http://www.cervantesvirtual.com/obra/guillem-de-varoich--0/

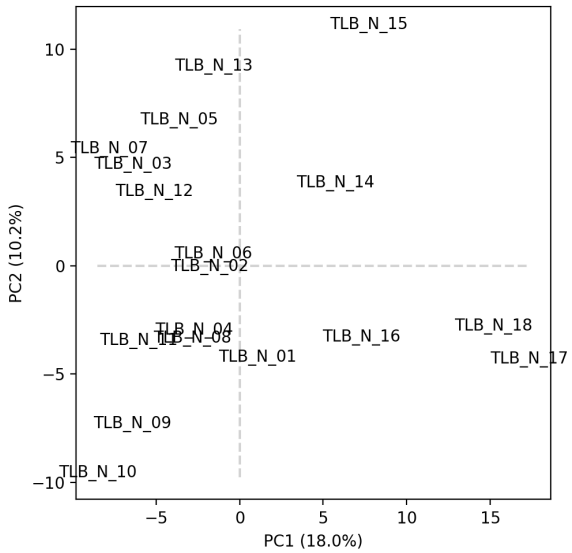

Figure 5: Scatter plot of PC1 and PC2 for the narration part with hyper-parameters $(n, r) = (3, 300)$.

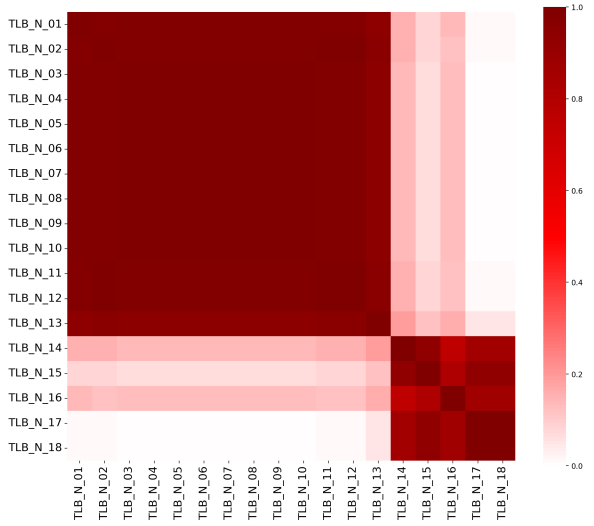

Figure 6: Pair-wise mean concordance rates for the narration part that was computed from 100 iterations of k-means performed with the hyper-parameters $(k, n, r) = (2, 3, 300)$. The darker the cell, the more similar the pair of pieces.

ternal clusters. Although this speculation requires verification by conducting experiments with undisputed works, we argue that if a different hand had come into play, both narration and conversation would coincide at the cluster boundary, which is not the case with $k = 2$. Nonetheless, when k is set to three for the conversation part, there emerges a subcluster within the second cluster, whereas the first cluster remains intact, as displayed in Figure 9. This subcluster comprises TLB_C_22–TLB_C_24, corresponding approximately to chapters 355–487. This boundary agrees well with that detected for the narration part at chapter 350, as noted above. In line with Martorell and de Galba (1947), the concurrence of the boundaries suggests that, if Galba had made some contribution to TLB, it should be located from chapter 350 to the end. The fact that new boundaries do not emerge when k is set to four or five points to strong internal cohesion of the clusters.

5.2 Detection of Number of Components

Thus far, it is evident that internal variation exists both in the narration and conversation parts. However, we are yet to verify if the variation is so large as to ascribe it to different authors. Despite the detection of the correct number of components being a challenging problem in stylometry (Koppel et al., 2011), we attempted to statistically determine the number of distinct hands that may have participated in TLB. We presumed that if single-authorship was more likely, only one component would be detected instead of two or more components, in which case multiple-authorship would be backed up.

By formulating the problem as model selection, we applied a Gaussian Mixture Model (GMM) combined with Bayes Information Criterion (BIC). GMM allows for probabilistic clustering to explore the heterogeneity in multivariate data (Frühwirth-Schnatter, 2006; Murphy, 2012). Combination with BIC enables model selection, considering the trade-off between model complexity and goodness of fit; a smaller BIC value indicates a better model. The capability of the algorithm to estimate the correct number of components has been demonstrated in the literature (Leroux, 1992). Although its effectiveness for stylometric studies requires empirical validation with the works of undisputed authorship, it will be worthwhile to apply the method to our case of interest. We implemented the algorithm using `sklearn.mixture.GaussianMixture` (Pedregosa et al., 2011)[8] with the default `full` covariance parameter and varying the number of components $k \in \{1, 2, 3, 4, 5\}$.

Figure 10 reveals that the effective number of components is $k = 1$ for every r in the narration. An identical pattern was observed for the conversation part. The behavior was consistent across the

[8] https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

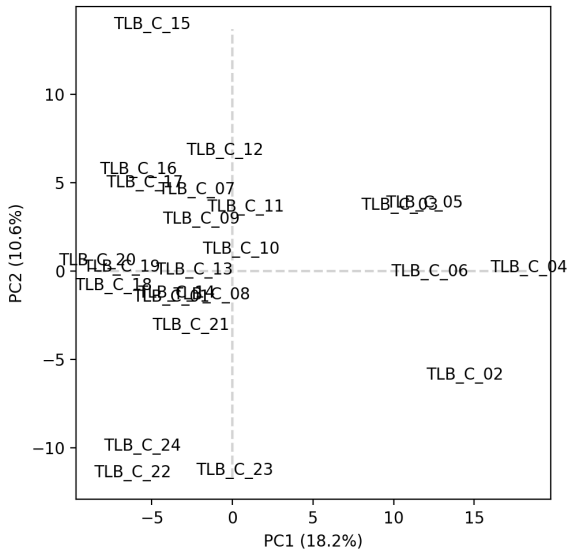

Figure 7: Scatter plot of PC1 and PC2 for the conversation part with hyper-parameters $(n, r) = (3, 300)$.

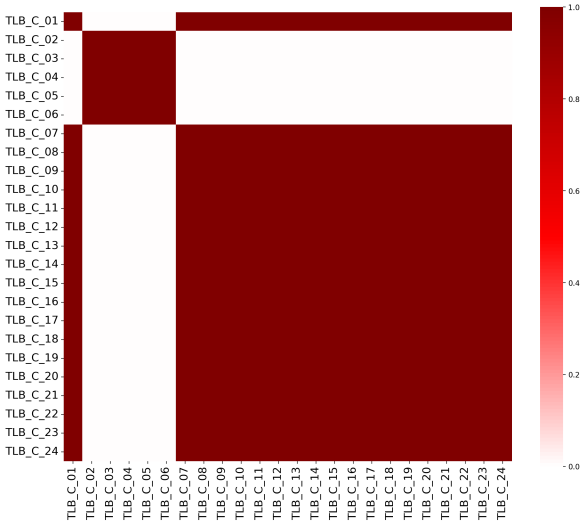

Figure 8: Pair-wise mean concordance rates for the conversation component that was computed from 100 iterations of k-means performed with the hyper-parameters $(k, n, r) = (2, 3, 300)$. The darker the cell, the more similar the pair of pieces.

hyper-parameter space (n, r) (figures not shown) except for $(n, r) = (1, 33)$, in which case the estimated number of components was two for narration and three for conversation. The fact that the outcome converges as n grows larger would justify giving more importance to the results obtained with $n \geq 2$. We suspect that unigrams are too coarse-grained to elicit an immanent pattern.

Consequently, we argue that, despite internal differences, single-authorship is more likely than dual-authorship from a statistical viewpoint. We conjecture that the clear split observed with PCA and k-means simply reflects Martorell's internal stylistic variation without necessarily pointing to different authorship. Alternatively, Galba might have actually contributed to the creation of the fourth part starting from around chapter 350 onward to the end, but too little for his own stylistic fingerprints to be recognized.

5.3 Distinctive POS *n*-grams

We explored POS n-grams that played a crucial role in the multivariate analyses and deserve special mention from the philological viewpoint. As we explained in Section 3, the relative frequencies of n-grams were standardized to have zero mean and unit variance for every variable. An n-gram was considered distinctive when its absolute value was above 1 on average for the pieces of interest.

With respect to the narration part, we focus on the pieces TLB_N_14–TLB_N_18 forming a cluster in Figure 6. In these pieces, the trigrams that include ADJ_N are frequently used: MOLT_**ADJ_N**, **ADJ_N**_I, and **ADJ_N**_ADV. The sequence ADJ_N represents the preposition (instead of posposition) of an adjective to the noun that it modifies (e.g., *triümphal victòria* "triumphant victory"). Coromines (1971) attributed the excessive use of epithet preposition to the alleged Galba's contribution. Also characteristic are the sentences beginning with the conjunction *i* "and" followed by a verb, as illustrated by **PUNCT_I_V** and **PUNCT_I_FER** (e.g., . *E lexaren* ". And they left"). Other distinctive features include the use of the adverb *molt* "very", as exemplified by **MOLT**_ADJ_V and **MOLT**_ADV_N (e.g., *molt bé acompanyats* "very well accompanied"), and that of the adjective *gran* "great", as seen in **GRAN**_N_I (e.g., *gran importància e* "great importance and").

In the conversation part, we first focus on the pieces TLB_C_02–TLB_C_06 forming a cluster in Figure 8. In these pieces, the trigrams that include ART_N representing a noun phrase headed by an article (e.g., *lo rey* "the king") are extensively used: PUNCT_**ART_N**, **ART_N**_V, **ART_N**_I, COM_**ART_N**, **ART_N**_ÉSSER, ADV_**ART_N**, DE_**ART_N**, **ART_N**_ADV, AB_**ART_N**, and V_**ART_N**. This usage reflects the abundance of narrational components within the conversation. In contrast, the following trigrams appear much fewer times: PRON_**HAVER_V**, which contains present perfect construction formed by *haver* "to have" and

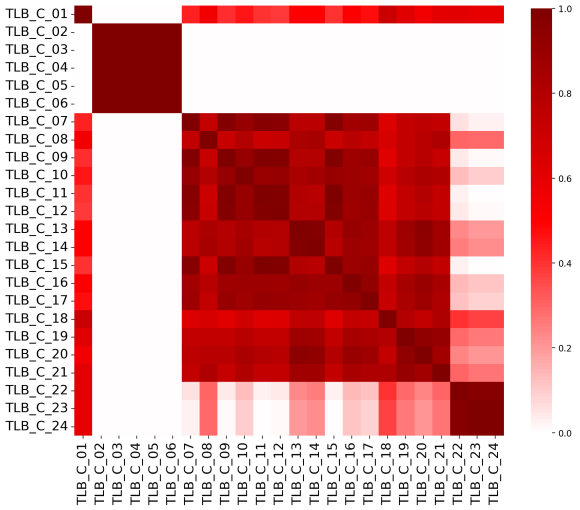

Figure 9: Pair-wise mean concordance rates for the conversation component that was computed from 100 iterations of k-means performed with the hyper-parameters $(k, n, r) = (3, 3, 300)$. The darker the cell, the more similar the pair of pieces.

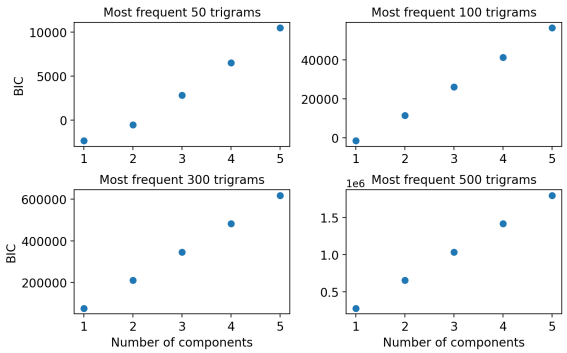

Figure 10: Model selection with Gaussian Mixture Model for narration part. The hyper-parameters are $n = 3$ and $r \in \{50, 100, 300, 500\}$. A smaller BIC value indicates a better model.

past participle (e.g., *·ns ha dats* "has given us"); V_**PUNCT**_**CONJ** and N_**PUNCT**_**CONJ**, which represent a sentence beginning with conjunction (e.g., *glòria. Donchs* "glory. So"); and **ADJ**_**N**_V, **ADJ**_**N**_CONJ, ART_**ADJ**_**N**, etc, all of which represent preposition of an adjective to the noun that it modifies.

The following trigrams characterize a subcluster TLB_C_22–TLB_C_24 in Figure 9 for frequent occurrences: **HAVER**_**V**_ART, which contains present perfect construction (e.g., *ha presa la* "has caught the"); **V**_**V**_I, which involves, for instance, an infinitive preceded by an auxiliary verb including *poder* "can" and *voler* "to want" (e.g., *podia veure e* "could see and"); and ART_**ADJ**_**N**, **ADJ**_**ADJ**_**N**, etc., all of which represent preposition of an adjective to the noun that it modifies. Its frequent use is also seen in the corresponding narration part (Coromines, 1971).

6 Conclusions and Future Work

This study revisited the unsettled authorship attribution of *Tirant lo Blanc* using stylometric techniques; specifically, we exploited POS n-grams as stylistic features. Furthermore, we addressed the distinction between narration and conversation, which has hitherto been omitted. We performed exploratory multivariate analyses and demonstrated that, despite internal differences, single-authorship is more likely from a statistical point of view. If Galba had contributed something to the last quarter of the work, it would have been minimal.

One limitation of our study is the adoption of rather coarse granularity in parts-of-speech. For instance, we did not distinguish between verbal forms such as finite forms, infinitive, gerund, and participle and instead treated them all under the category of VERB. However, their peculiar usage has been pointed out in previous literature (Gili i Gaya, 1947; Ferrando, 1987; de Riquer, 1990; Ferrando, 2012) and so could be useful for detecting authorial fingerprints as well. Furthermore, it will be intriguing to explore the orthographic, lexical, morphological, and syntactic traits that have been suggested as distinctive in previous research (Gili i Gaya, 1947; Coromines, 1971; Ferrando, 1987; Skubic, 1989; de Riquer, 1990; Ferrando, 2012), to name a few[9].

Moreover, two hypotheses concerning the genesis of TLB remain to be examined stylometrically: (i) that a short fragmentary manuscript denominated *Guillem de Vàroich* was actually written by Martorell (Gili i Gaya, 1947; de Riquer, 1990); and (ii) that the Valencian writer Joan Roís de Corella is the genuine author of TLB (Guia i Marín, 1996).

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP18K12361 and JP23K12152.

References

Harald Baayen, Hans van Halteren, and Fiona Tweedie. 1996. Outside the Cave of Shadows: Using Syntac-

[9]See also: http://www.cervantesvirtual.com/portales/joanot_martorell_i_el_tirant_lo_blanc/llengua/

tic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 11(3):121–132.

John Burrows. 2002. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Miguel de Cervantes Saavedra. 1999. El Ingenioso Hidalgo Don Quijote de la Mancha.

Ross Clement and David Sharp. 2003. Ngram and Bayesian Classification of Documents for Topic and Authorship. *Literary and Linguistic Computing*, 18(4):423–447.

Joan Coromines. 1971. Sobre l'estil i manera de Martí J. de Galba i el de Joanot Martorell. In *Lleures i converses d'un filòleg*, pages 363–378. Club Editor, Barcelona.

Maciej Eder. 2015. Does Size Matter? Authorship Attribution, Small Samples, Big Problem. *Digital Scholarship in the Humanities*, 30(2):167–182.

Antoni Ferrando. 1987. Entorn de la llengua del *Tirant lo Blanc*. *Estudis Romànics*, 4:369–372.

Antoni Ferrando. 2012. Llengua i context cultural al *Tirant lo Blanc*. Sobre la identitat del darrer Joanot Martorell (1458-1465). *eHumanista*, 22:623–668.

Marti Font, Xavier Puig, and Josep Ginebra. 2016. Bayesian Analysis of the Heterogeneity of Literary Style. *Revista Colombiana de Estadística*, 39(2):205–227.

Sylvia Frühwirth-Schnatter. 2006. *Finite Mixture and Markov Switching Models*. Springer, New York, NY.

Samuel Gili i Gaya. 1947. Noves recerques sobre *Tirant lo Blanch*. *Estudis Romànics*, 1:135–147.

Javier Girón, Josep Ginebra, and Alex Riba. 2005. Bayesian Analysis of a Multinomial Sequence and Homogeneity of Literary Style. *American Statistician*, 59(1).

Josep Guia i Marín. 1996. *De Martorell a Corella. Descobrint l'autor del Tirant lo Blanc*. Editorial Afers, Barcelona.

Graeme Hirst and Ol'ga Feiguina. 2007. Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22(4):405–417.

David I. Holmes. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3):111–117.

Patrick Juola. 2006. Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.

Yoshifumi Kawasaki. 2021. Stylometric Analysis of Avellaneda's *Don Quijote*. In *12th International Conference on Corpus Linguistics*, Universidad de Murcia (Online). Spanish Association for Corpus Linguistics.

Yoshifumi Kawasaki. 2022. A Stylometric Analysis of *Amadís de Gaula* and *Sergas de Esplandián*. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 1–7. Association for Computational Linguistics.

Mirco Kocher and Jacques Savoy. 2019. Evaluation of Text Representation Schemes and Distance Measures for Authorship Linking. *Digital Scholarship in the Humanities*, 34(1):189–207.

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1356–1364, Portland, Oregon, USA. Association for Computational Linguistics.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412.

Brian. G. Leroux. 1992. Consistent Estimation of a Mixing Distribution. *The Annals of Statistics*, pages 1350–1360.

Joanot Martorell. 2006. *Tirant lo Blanc*. Alacant : Biblioteca Virtual Joan Lluís Vives, 2006.

Joanot Martorell. 2008. *Tirant lo Blanch*. Tirant lo Blanch, València.

Joanot Martorell. 2016. *Tirant lo Blanc*. Labutxaca, Barcelona.

Joanot Martorell and Martí Joan de Galba. 1947. *Tirant lo Blanc: Text, introducció, notes i índexs*. Editorial Selecta, Barcelona.

T. C. Mendenhall. 1887. The Characteristic Curves of Composition. *Science*, 9(214S):237–246.

Kevin P Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, MA.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Yao Jean Marc Pokou, Philippe Fournier-Viger, and Chadia Moghrabi. 2016. Authorship Attribution Using Variable Length Part-of-Speech Patterns. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence*, volume 2, pages 354–361.

Xavier Puig, Martí Font, and Josep Ginebra. 2015. Classification of Literary Style that takes Order into Consideration. *Journal of Quantitative Linguistics*, 22(3).

Alex Riba and Josep Ginebra. 2005. Change-point estimation in a multinomial sequence and homogeneity of literary style. *Journal of Applied Statistics*, 32(1).

Alex Riba and Josep Ginebra. 2006. Diversity of Vocabulary and Homogeneity of Literary Style. *Journal of Applied Statistics*, 33(7).

Martí de Riquer. 1990. *Aproximació al Tirant Lo Blanc*. Quaderns Crema, Barcelona.

Jacques Savoy. 2017. Analysis of the Style and the Rhetoric of the American Presidents over Two Centuries. *Glottometrics*, 38:55–76.

Mitja Skubic. 1989. L'estructuració de l'oració composta en el *Tirant lo Blanc*. *Linguistica*, 29(1):137–145.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Ayaka Uesaka and Masakatsu Murakami. 2015. Verifying the authorship of Saikaku Ihara's work in early modern Japanese literature; A quantitative approach. *Digital Scholarship in the Humanities*, 30(4):599–607.

Mario Vargas Llosa. 2015. *Carta de batalla por Tirant lo Blanc*. Debolsillo.

C. B. Williams. 1975. Mendenhall's Studies of Word-Length Distribution in the Works of Shakespeare and Bacon. *Biometrika*, 62(1):207–212.