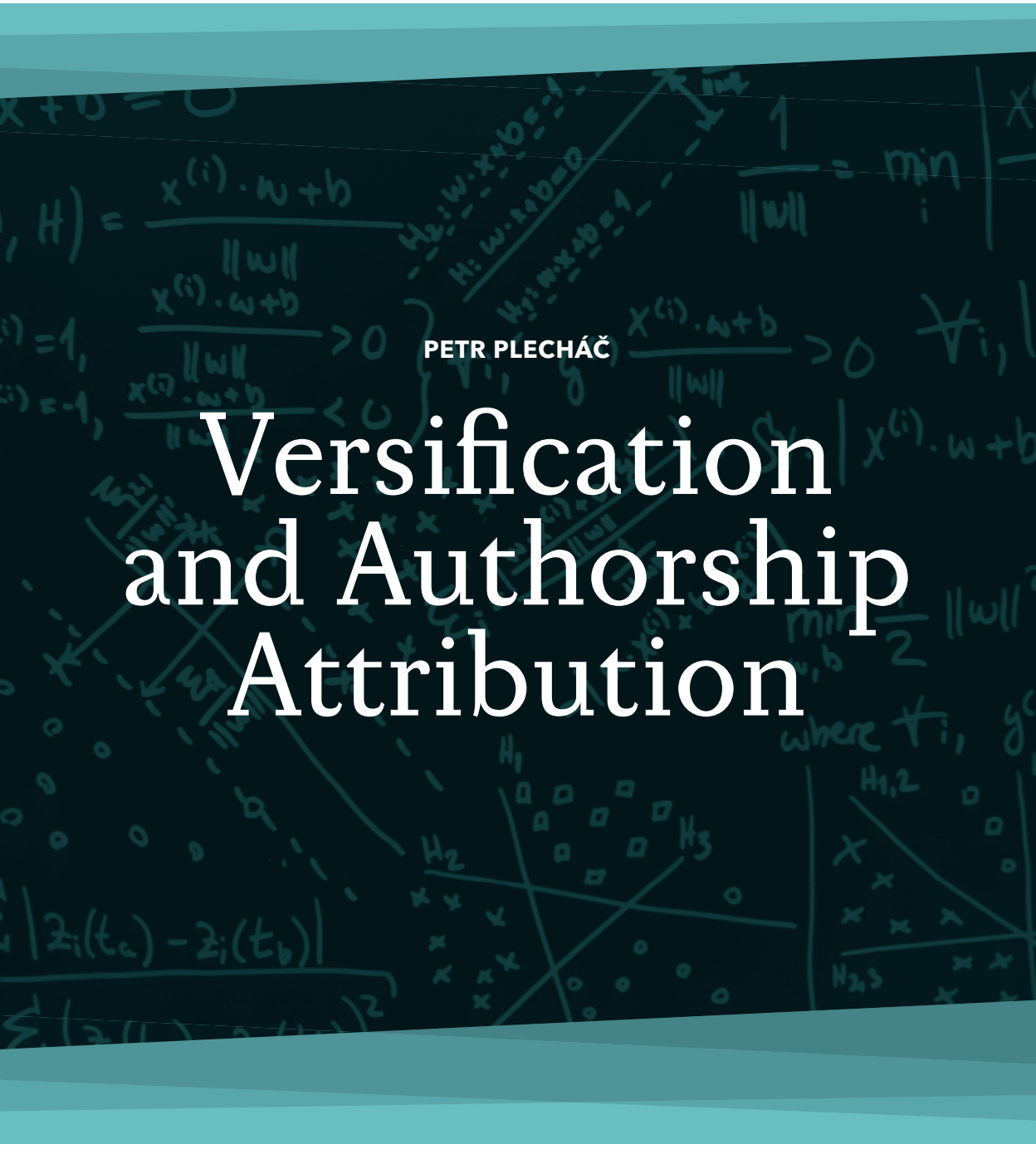
CZECH LITERATURE STUDIES



PETR PLECHÁČ

Versification and Authorship Attribution

INSTITUTE OF CZECH LITERATURE
KAROLINUM PRESS

4 Applications

Petr Plecháč
Institute of Czech Literature, Czech Academy of Sciences
e-mail: plechac@ucl.cas.cz

Artjoms Šeļa
Institute of Polish Language, Polish Academy of Sciences / University of Tartu
artjoms.sela@ijp.pan.pl

https://doi.org/10.14712/9788024648903.5

4 Applications

In this final chapter, I apply the approaches described in this book to two cases of ambiguous or disputed authorship of poetic works. These cases concern English and Russian texts respectively. In the first, I trace the relative contributions of William Shakespeare and John Fletcher to the play *The Two Noble Kinsmen.* Then, in the second, I collaborate with Artjoms Šeļa to investigate the suspected forgery of poems first published in a 1978 edition of Gavriil Batenkov's works.

4.1 *The Two Noble Kinsmen*

The play *The Two Noble Kinsmen* (*TNK*) was recorded in the Stationers' Register in 1634 and published in a quarto edition later that year. In both cases, John Fletcher and William Shakespeare were indicated as the play's authors. No manuscript has been preserved. Attempts to discern which parts were likely written by each author must therefore rely solely on intratextual indicators. Since the 19th century, researchers have found evidence at various textual levels to suggest that Shakespeare was mostly responsible for Acts 1 and 5 while Fletcher was mostly responsible for Acts 2, 3 and 4.[20] While there is not much controversy about this general picture, the authorship of certain scenes is still being debated. In what follows, I seek to contribute to this debate using a combination of versification-based and word-based models.

The case of *TNK* is closely linked to that of another play which was also supposedly co-authored by Shakespeare and Fletcher—*The Famous History of the Life of King Henry the Eight*. I have discussed the authorship of that work elsewhere (Plecháč 2020). Here I follow the design of that study and apply the same models to classify passages from *TNK*.

20 A detailed history of *TNK*'s attributions is given in Vickers 2004.

https://doi.org/10.14712/9788024648903.5

				I						II		
	P	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**6**
Weber 1812	N	S	S	S	S	S	F	F	F	F	F	F
Spalding 1833	N	S	S	S	S	S	F	F	F	F	F	F
Hickson 1847	N	S	SF	S	S	S	S	F	F	F	F	F
Fleay 1874	N	S	S	S	S	S	S	F	F	F	F	F
Boyle 1882	N	M	M	M	M	M	M	F	F	F	F	F
Oliphant 1891	N	FSM	SM	SM	SM	?	S	F	F	F	F	F
Farnham 1916	N	S	S	S	S	S	F	F	N	F	F	F
Hart 1934	N	S	S	S	S	S	F	F	F	F	F	F
Oras 1953	N	S	S	S	N	N	N	F	F	F	F	F
Hoy 1962	N	S	S	S	S	S	S	F	F	F	F	F
Horton 1987	N	S	S	S	N	N	S	F	S	N	F	N
Matthews-Merriam 1993	N			S						F		
Ledger-Merriam 1994	F	S	S	S	S	?	S	F	F	F	F	?
Tarlinskaja 2014	N	S	S	S				F	F	F	F	F
Eisen et al. 2017	N	S	S	S	S	F	S	F	?	F	F	F

TAB. 4.1: Selected attributions of *TNK*. *S* denotes an attribution of the scene to Shakespeare, *F* to Fletcher and *M* to Massinger; *N* denotes an unassigned scene.

4.1.1 History and Related Works

The first attempt to provide a scene-by-scene division of *TNK* between Shakespeare and Fletcher was made by Henry Weber (1812). Based on his observations of enjambments, weak endings, unusual words and metaphors, Weber assigned all of Act 1 and most of Act 5 to Shakespeare and all of Act 2 and most of Acts 3 and 4 to Fletcher (see TAB. 4.1 for details of this and other attributions). Slightly different attributions were proposed by William Spalding (1833) and Samuel Hickson (1847), both of whom relied on observations similar to those of Weber.

An important advance came with the publication in the 1874 *Transactions of the New Shakspere Society* of three articles about the play which instead of merely observing distinctive features sought to quantify them: Frederick Gard Fleay (1874d) measured the number of weak endings and four-feet lines in particular scenes; Frederick James Furnivall (1874c) considered the number of enjambments (the stopt-line test); and John Kells Ingram (1874) applied his weak-ending test (see Section 1.1). All three articles supported Hickson's division with only one exception—Act 1, scene 2 was now assigned solely to Shakespeare.

Just a few years later, Robert Boyle (1882) presented a new theory which claimed that the "Shakespearian" parts had in fact been written by Philip Massinger or—in two cases—by

III						IV			V				
1	2	3	4	5	6	1	2	3	1	2	3	4	E
S	S	F	F	F	F	F	F	S	S	F	S	S	N
S	F	F	F	F	F	F	F	F	S	F	S	S	N
S	S	F	F	F	F	F	F	S	S	F	S	S	N
S	S	F	F	F	F	F	F	S	S	F	S	S	N
SM	SM	F	F	F	F	F	F	M	M	F	M	M	N
S	S	F	F	F	F	F	F	FS	FS	F	S	S	F
S	?	F	F	F	F	F	F	S	?	F	S	S	N
S	F	F	F	F	F	F	F	F	S	F	S	S	N
S	?	F	F	F	F	F	F	S	S	F	S	S	N
S	S	F	F	F	F	F	F	F	FS	F	S	S	N
S	N	F	N	?	F	?	?	S	S	?	S	S	N
			F				S				S		N
S	?	S	?	F	F	S	S	S	S	F	S	S	?
S	F	F	F	F	F	F	S		S	F	S	S	N
S	S	F	F	F	F	F	F	S	S	N	S	S	N

Shakespeare and Massinger together. Massinger's participation was also backed by Henry Oliphant (1891) although he pointed to different scenes to those named by Boyle.

Twentieth-century studies generally supported the Shakespeare–Fletcher division that preceded Boyle or else proposed only slight modifications. These works included studies of contractions (Farnham 1916), vocabulary richness (Hart 1934), line endings (Oras 1953) and spelling differences (Hoy 1962).

This Shakespeare–Fletcher split has also largely been maintained by more recent scholars. Based on a discriminant analysis of three sets of function words, Thomas Horton (1987) attributed most scenes in the play to Shakespeare or else left them undecided. Robert Matthews and Thomas Merriam (1993) classified entire acts of *TNK* using a neural network that had been familiarised with the frequencies of function words in the respective plays of Shakespeare and Fletcher. A year later, Merriam reopened the case in a study with Gerard Ledger which used a hierarchical cluster analysis based on character frequencies; this time the goal was the attribution of particular scenes (Ledger and Merriam 1994). More recently, Marina Tarlinskaja (2014) has applied a complex versification analysis using features of the kind enumerated in Section 1.5. Mark Eisen, Alejandro Riberio, Santiago Segarra and Gabriel Egan (2017) have also used word adjacency networks (Segarra, Eisen and Riberio 2013) to analyse the frequencies of collocations of selected function words in particular scenes of the play.

4.1.2 Attribution of Particular Scenes

Since the external evidence clearly pointed to Shakespeare and Fletcher's joint authorship of *TNK* and previous analyses had ruled out Massinger's participation on linguistic grounds, I limited the candidate set to Shakespeare and Fletcher. I then set out to determine the most likely author of particular scenes.

To train the models, I used four plays by Shakespeare and four plays by Fletcher that all dated roughly from the period when *TNK* was supposedly written (1613–1614). Each scene in these plays was treated as a single training sample except for those containing fewer than 100 verse lines. This gave me:

— Shakespeare: *The Tragedy of Coriolanus* (5 scenes), *The Tragedy of Cymbeline* (10 scenes), *The Winter's Tale* (7 scenes), *The Tempest* (6 scenes) and
— Fletcher: *Valentinian* (12 scenes), *Monsieur Thomas* (10 scenes), *The Woman's Prize* (14 scenes), *Bonduca* (14 scenes).[21]

Altogether there were, thus, 28 training samples for Shakespeare and 50 training samples for Fletcher.

Having established a large enough training set, I now risked employing a method that might produce rather sparse data: First I used the frequencies of particular rhythmic types to capture the rhythmic style of the data (cf. Section 2.1.2).[22] No rhyme characteristics were considered since all of the plays were written in blank verse and rhymes were, thus, only exceptional. To capture vocabulary, I relied on word frequencies since words had proven to be a more reliable indicator than lemmata at the pilot testing stage. For both rhythmic types and words, I limited the analysis to the 500 most frequent types. An SVM with a linear kernel was used as a classifier.

To estimate the model's accuracy, I performed the following cross-validation:

— To avoid overfitting—a potential risk of testing a model on scenes from the play it was trained with—I did not perform standard *k*-fold cross-validation. Instead, I classified scenes from each play using a model trained with the rest of the plays. As such, scenes from Shakespeare's *Coriolanus* were classified by a model

21 For both the training data and the text of *TNK* itself, I relied on XML versions of the first editions of the plays, as provided by the EarlyPrint project (https://drama.earlyprint.org). To eliminate spelling variation, regularised spellings (the "reg" attribute of the w-element) were used. All of Shakespeare's texts came from the First Folio (1623). All of Fletcher's texts came from the first Beaumont and Fletcher folio (1647), except for *Monsieur Thomas* for which the 1639 quarto was used. For *TNK*, I relied on the 1634 quarto edition.

22 Rhythmic annotation was provided by the Prosodic Python library (https://github.com/quadrismegistus/prosodic).

		Rhyt. type-based models	Word-based models	Combination models
Shakespeare	*Coriolanus*	1	1	1
	Cymbeline	0.997	1	1
	The Winter's Tale	1	1	1
	The Tempest	1	1	1
	Valentinian	0.992	1	1
Fletcher	*Monsieur Thomas*	1	1	1
	The Woman's Prize	1	1	1
	Bonduca	1	0.93	1

TAB. 4.2: Accuracy of authorship recognition by models based on (1) the 500 most common rhythmic types, (2) the 500 most common words and (3) 1000-dimensional vectors combining features (1) and (2). Figures show the share of correctly classified scenes over all 30 iterations.

trained with scenes from the other three plays by Shakespeare and four plays by Fletcher; 27 scenes from *Cymbeline* were classified similarly and so on.

— Since the training data were imbalanced and there was, thus, a risk of bias, I aligned the number of training samples per author using random selection.

— To obtain more representative results, the entire process was repeated 30 times with a new random selection in each iteration; this generated 30 classifications of each scene.

— To compare the attribution power of both feature subsets, cross-validation was performed not only on the combined models (500 rhythmic types ∪ 500 words) but also on the versification-based models (500 rhythmic types) and word-based models (500 words) alone.

As TAB. 4.2 shows, both versification-based and word-based models proved highly accurate in distinguishing the respective works of Shakespeare and Fletcher. The only issues with the versification-based models were one misattribution of Act 3, scene 5 of *Cymbeline* to Fletcher and two misattributions of Act 5, scene 8 of *Valentinian* to Shakespeare. In contrast, the word-based models misclassified Act 5, scene 1 of *Bonduca* in all 30 iterations. When the two feature sets were merged, however, there were no misclassifications and all models achieved 100% accuracy.

FIG. 4.1 presents the results of the application of classifiers to *TNK*. As with the training samples, testing was limited to scenes with more than 100 lines (12 out of the play's 24 scenes). Except in the case of Act 4, scene 1, there was a strong consensus among the versification-based, word-based and combined models. Significantly, their

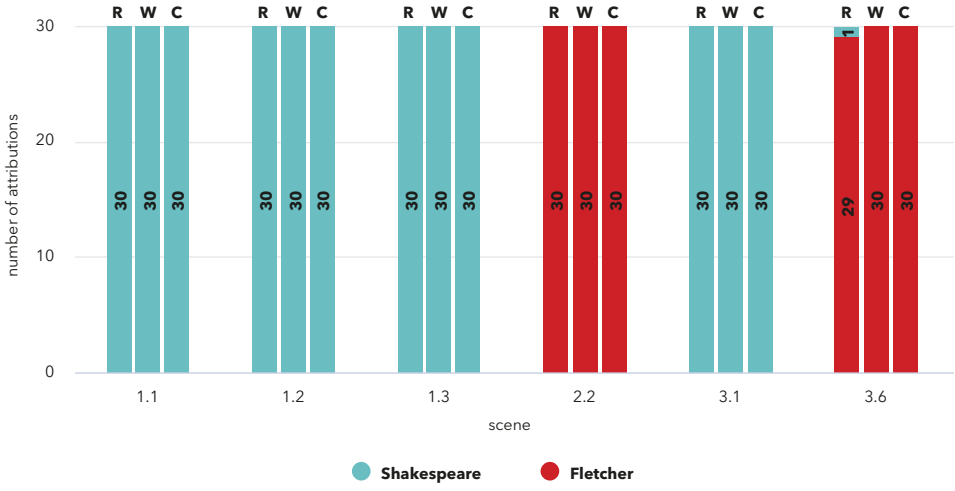


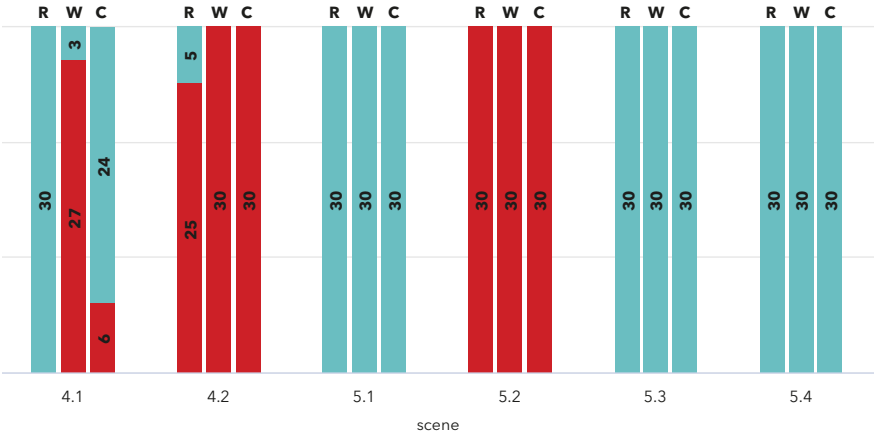
FIG. 4.1: Classification of *TNK* scenes with more than 100 lines by versification-based models (R), word-based models (W) and combined models (C). The figure shows the number of times per 30 iterations that the author was credited with a given scene.

predictions also reflected the attributions of scholars such as Fleay (1874d) and Oras (1953). Concerning Act 4, scene 1, there were mixed signals. Versification-based models unanimously assigned the scene to Shakespeare, but almost all the word-based models attributed it to Fletcher. The combined models again favoured Shakespeare.

This classification of particular scenes may have been strong evidence of the involvement of both authors. Nevertheless, since only half of *TNK*'s scenes were long enough to be classified, this approach did not allow me to estimate the overall contributions of each author. To trace authorial signals through all the versified parts of the play, I therefore proceeded with a different technique. This was *rolling attribution*, a method originally proposed by Maciej Eder (2016).

4.1.3 Rolling Attribution of *TNK*

The logic behind the rolling approach was quite simple. Instead of classifying particular scenes from *TNK* with a model trained with complete scenes from different plays, the plays in the training set were split into 100-line samples that disregarded scene divisions. Here sample 1 was lines 1–100 of the play; sample 2 was lines 101–200;



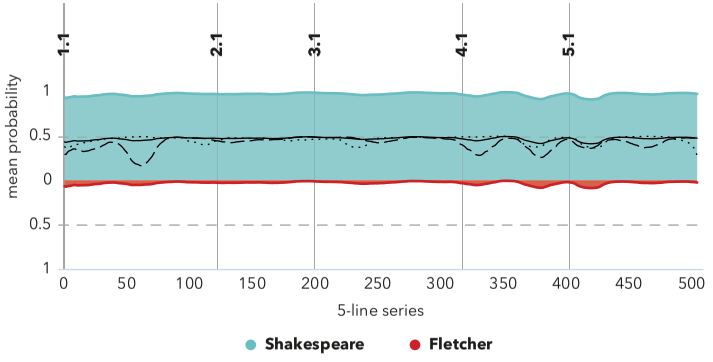
sample 3 was lines 201–300 and so on. An SVM model trained with these samples was then used to classify 100-line samples from *TNK*. To trace potential authorship shifts more precisely, the *TNK* samples were not extracted successively as the training data had been. Instead, a "rolling" window of 100 lines was established and set to advance in five-line steps (thus, sample 1: lines 1–100, sample 2: lines 6–105, sample 3: lines 11–110 and so on).

This rolling attribution scheme was first tested with the plays contained in the training set. For each play, I trained 30 models with the remaining data, having aligned the number of training samples by random selection in each iteration. To enhance authorship detection even further, I avoided binary classification (author = Shakespeare | author = Fletcher) and instead transformed the output into a probability distribution between the two authors via Platt scaling (Platt 1999).

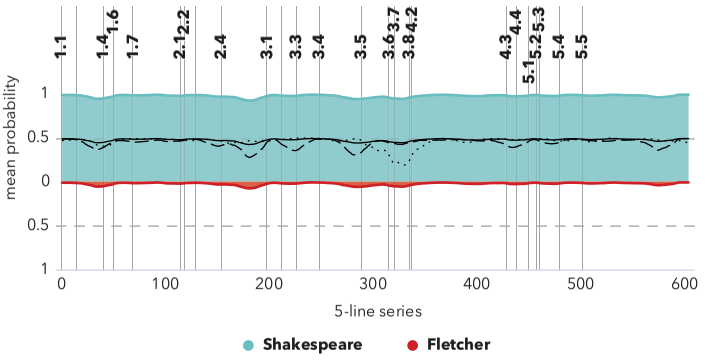
I focused here not on the complete samples but rather on the successive series of five lines. With a sample size of 100 lines, a "step" set to five lines and 30 different models, each five-line series in *TNK* (except for the initial 19 and final 19 series) was classified 600 times—that is, 30 times within 20 different samples. I averaged out the probabilities obtained from the different models and samples for these series.

FIG. 4.2 shows the results for the combined models as well as those for the versification-based models and word-based models on their own. The versification-based models produced several misclassifications. In particular, 15 series from Act 4, scene 1 and two series from Act 5, scene 8 of *Valentinian* were misattributed to Shakespeare. The probabilities of Shakespearean and Fletcherian authorship also came close in a couple of series in Act 2 scene 1 of *Bonduca* although Shakespeare's values remained slightly higher. Nevertheless, since there were only 17 misclassifications out of a total

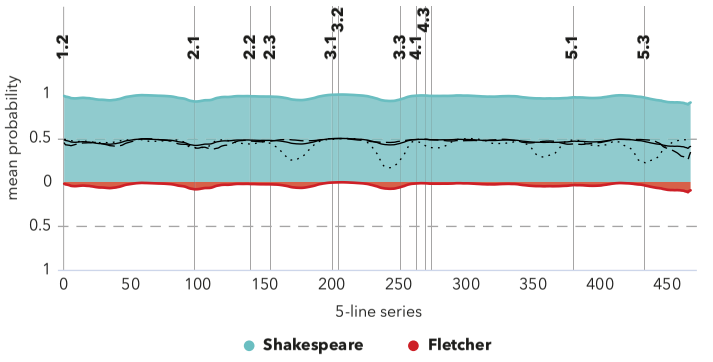
FIG. 4.2: Rolling attribution of four plays by Shakespeare and four plays by Fletcher based on the 500 most common rhythmic types and the 500 most common words (upper = *P*(Shakespeare), lower = *P*(Fletcher)). Vertical lines indicate scene breaks. Dotted lines show the results of rolling attribution based solely on the 500 most common rhythmic types. Dashed lines show the results of rolling attribution based solely on the 500 most common words.



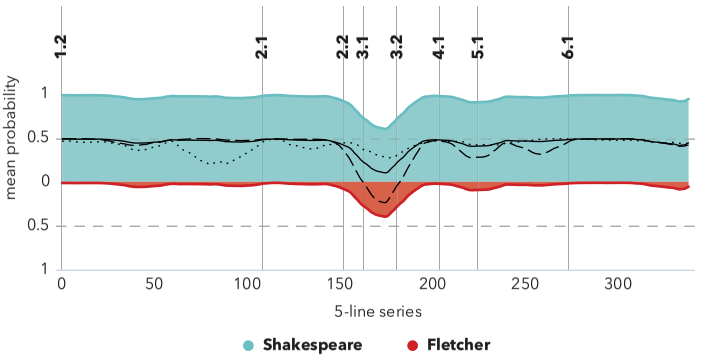
(a) Shakespeare: *Coriolanus*



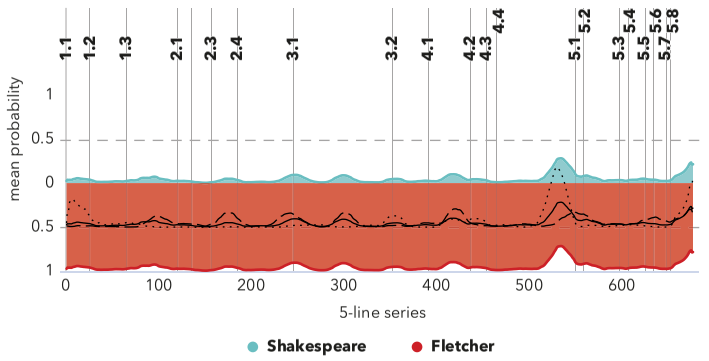
(b) Shakespeare: *Cymbeline*



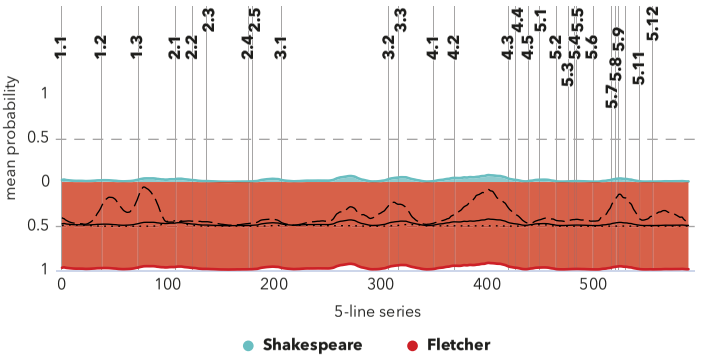
(c) Shakespeare: *The Winter's Tale*



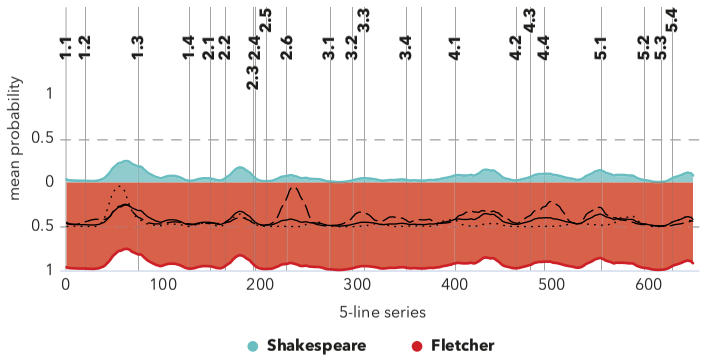
(d) Shakespeare: *The Tempest*



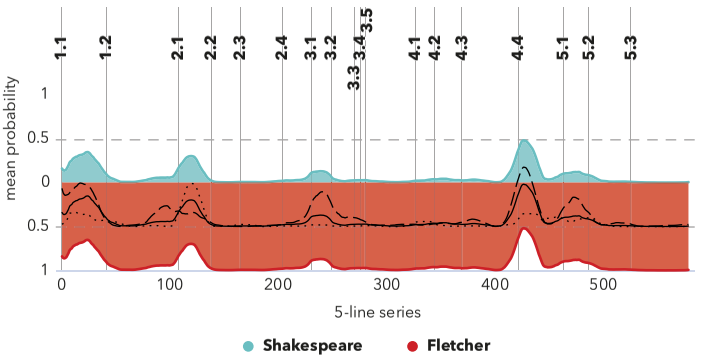
(e) Fletcher: *Valentinian*



(f) Fletcher: *Monsieur Thomas*



(g) Fletcher: *The Woman's Prize*



(h) Fletcher: *Bonduca*

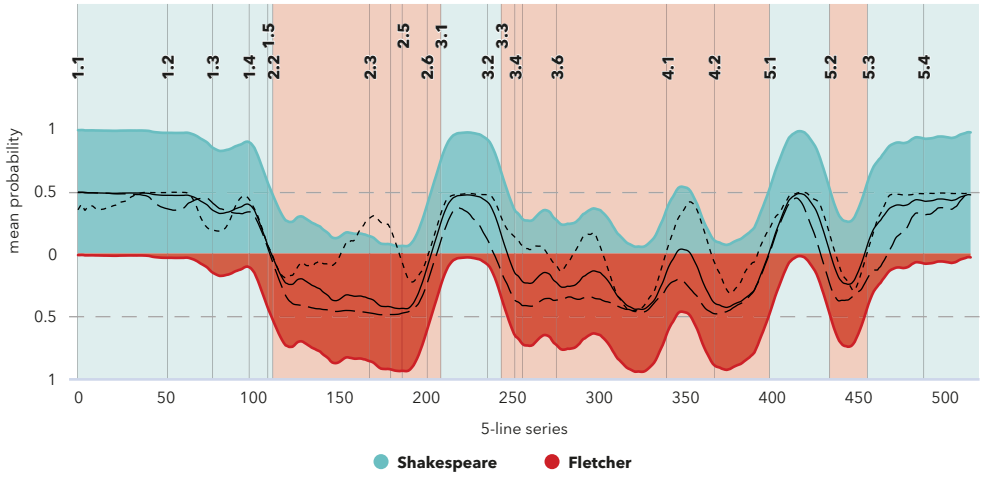


FIG. 4.3: Rolling attribution of *TNK* based on the 500 most common rhythmic types and the 500 most common words (upper = *P*(Shakespeare), lower = *P*(Fletcher)). Vertical lines indicate scene breaks. Dotted lines show the results of rolling attribution based solely on the 500 most common rhythmic types. Dashed lines show the results of rolling attribution based solely on the 500 most common words. The background colour indicates the author to whom the scene is usually credited.

4412 series, the overall accuracy rate was high at 0.996. Word-based models also gave rise to misclassifications: in Act 3, scene 1 of *The Tempest*, 20 series were misattributed to Fletcher while 14 series in Act 4, scene 4 of *Bonduca* were wrongly assigned to Shakespeare; Shakespeare and Fletcher were again weighted similarly for Act 1, scene 1 of *Bonduca*. Total accuracy was, thus, 0.992.

Crucially, all of these outlying results were absorbed and no series was misclassified when the feature sets were merged in the combined model.

Having verified the performance of the models, I turned to the evaluation of *TNK* itself. FIG. 4.3 gives the results of the rolling attribution of the play using models trained with all eight plays in the training set.

There were some remarkable discrepancies between the results of the versification-based and word-based models. This applied especially to the following sequences of *TNK*: from the end of Act 2, scene 2 to the end of Act 2, scene 4; from Act 3, scene 3 to Act 3, scene 5; and during Act 3, scene 6 and Act 4, scene 1. Interestingly enough, these were not controversial scenes where such behaviour might

be expected but rather parts of play whose attribution had been stable for the last two centuries (cf. TAB. 4.1). We may, however, be guided here by the combined model, which had proven most reliable and produced somewhat consistent results with *TNK* as well.

Based on these findings, it was highly probable that Shakespeare was the author of all of Act 1 and Fletcher was the author of all of Act 2. Indeed, the shift in authorship seemed to coincide with the break between these examined parts. (Significantly, Act 2, scene 1, which is usually assigned to Shakespeare, was excluded from the data because it was written in prose.) The models strongly favoured Shakespeare again in Act 3, scenes 1 and 2 (or, more precisely, from the end of Act 2, scene 6 to the start of Act 3, scene 3). For the remainder of Act 3 and Act 4 (excluding the prose text of Act 4, scene 3), Fletcher was the preference for all but nine series in Act 4, scene 1. As the final act opened, the likelihood of Shakespearean authorship rose sharply again and it remained high until the end of the play except in Act 5, scene 2 where Fletcher was the clear choice of the models. Again the authorial changes seemed to match scene breaks precisely.

All in all, then, the models strongly supported Shakespeare as the author of Act 1, scenes 1–5; Act 3, scenes 1–2; and Act 5, scenes 1 and 3–4. Similarly, they backed Fletcher as the author of Act 2, scenes 2–6; Act 3, scenes 3–6; Act 4, scene 2; and Act 5, scene 2. The authorship of Act 4, scene 1 remained uncertain. Notably, these results confirmed the attributions proposed by Fleay (1874d) and Oras (1953). Given that these scholars and others have provided (mostly orthogonal) evidence for Fletcher's authorship of Act 4, scene 1, it is tempting to lean towards the same conclusion.

4.1.4 Summary

Combined versification- and word-based models turned out to be highly accurate in distinguishing the work of Shakespeare from that of Fletcher. In the case of *The Two Noble Kinsmen,* the application of these models to particular scenes—especially when paired with rolling attribution—supported what may be called the orthodox division of the play.

These findings clearly testify to the efforts of the brilliant scholars who were able decades or more ago to identify the most salient features of the two authors' styles without the aid of any machines or feature selection algorithms. Instead they relied solely on thorough study of the texts in question. The features they pinpointed—for instance, the frequencies of *'em / them* (Thorndike 1901), *th'*, *i'* (Farnham 1916) and *doth* and *ye* (Hoy 1962)—all rank among those found to be most important for the

classification.[23] This is also true of line endings (Fleay 1874d). Common strong ending rhythmic types such as 0101010101 (0 = unstressed syllable, 1 = stressed syllable), 0101000101 and 0100010101 were among the most strongly-weighted positive (Shakespearian) features. Similarly, common W-position-terminated rhythmic types such as 01010101010 and 10010100010 appeared among the most strongly-weighted negative (Fletcherian) features.

4.2 The Case of (Pseudo-)Batenkov: Towards a Formal Proof of Literary Forgery *(co-authored by Artjoms Šeļa)*

In 1978, a scholarly monograph about the poetry of G. S. Batenkov (1793–1863) was published in Moscow under the title *Poezia dekabrista Gavriila Stepanovicha Batenkova* (Iliushin 1978). Its author was A. A. Iliushin. What appeared to be a complete collection of Batenkov's poems was appended to the volume.

Batenkov, a Russian officer and poet, had fought in the Napoleonic wars and later worked as an engineer and policymaker. His eclectic ideological interests, which ranged from freemasonry and Christian mysticism to political reform, led him to join secret societies and eventually become associated with the Decembrist revolt of 1825. This effectively ended his life as a free citizen of the Empire. He was sentenced to 25 years of solitary confinement in the Peter and Paul Fortress in Saint Petersburg and, after serving 20 years, exiled to Siberia.

Iliushin, who was both the author of the monograph and the editor of the appendix, was a Russian versification scholar and poetry specialist. He also wrote poetry himself and was known in academic circles for his literary games and imitations. The majority of Batenkov's late poems (i.e. those written after his release from prison) appeared for the very first time in this collection. There was, however, one major problem: the source of these texts was inaccessible and their origins unverifiable. Iliushin himself referred to a manuscript that was listed as lost in the archives (Shapir 2000).

For 20 years, no one publicly questioned the authenticity of these poems. This all changed when the scholar M. I. Shapir published a series of studies in the late 1990s that showed that there were indeed grounds for doubt. Shapir (1997, 1998) conducted

23 This appraisal is based on the mean value for feature importance in 30 combined models trained with 100-line samples taken from the training set (four plays by Shakespeare, four plays by Fletcher).

an extensive quantitative analysis of the poems in the controversial section of Batenkov's work (we refer to these texts as the "disputed poems"). To this end, he meticulously examined every linguistic level—prosody, metrics, morphology, syntax and semantics—and pointed out many significant differences between these texts and Batenkov's known works. Among the issues Shapir observed in the disputed poems were their abundance of inexact rhymes, overly archaic morphology, discrepancies in the use of pronouns and conjunctions and some possible anachronisms. To date, his analysis remains one of the most impressive non-computational authorship attribution studies of Russian poetry.

This research convinced many scholars that the disputed poems were in fact forgeries (Gasparov and Tarlinskaja 2008; Tarlinskaja 2014). Indeed, in the years since, this consensus has become so strong that the editors of a recently published collection of Iliushin's original poems did not hesitate to include all of the disputed poems in the volume (Iliushin 2020). However, this interpretation is at odds with Shapir's own conclusion: having uncovered significant differences at some textual levels but striking similarities at others, he judged that there was not enough evidence to draw any conclusions about the origins of the disputed poems. This reasoning led Shapir to an important generalisation about the limitations of using formal and linguistic methods to determine authorship. If, as he argued, we cannot trace the identity of an author based on various levels of linguistic features, then the concept of the "author" who makes linguistic choices that are unique and recognizable is nothing more than a scholarly construct.

From a modern-day perspective, Shapir's strong statements lack methodological support. Compared with other scholars who have used versification features for authorship attribution (Tomashevsky 1923/2008; Lotman and Lotman 1986; Tarlinskaja 2014), Shapir dramatically increased the number of textual levels under investigation. Nevertheless, his analysis remained univariate: all of the levels were treated in isolation and the features were compared one by one. It might be said, then, that Shapir's inquiry was multivariate in scope but he lacked the tools to deal with multivariate and seemingly contradictory signals. As a result, he could not estimate the compound authorial signal in either Batenkov's known works or the disputed poems. Key questions went unaddressed: How important were the differences in the frequency of inexact rhymes or function words compared, say, with similarities in the rhythmic structure of iambic tetrameter and use of formulae?

In the final part of this book, we return to this question that Shapir left unsolved. Our aim is to reach a more definitive conclusion about the authorship of the disputed poems using a multivariate approach that combines lexical and versification features. We break the problem down into the following experiments:

— We first test the general performance of our approach using 19th-century Russian poetry data.
— We then formulate the task as a *verification* problem. The goal here is not to find the most probable candidate from a finite set but rather to *verify* the likelihood that Batenkov's poems and the disputed poems were produced by a single author.
— Finally, we compare the disputed poems not only to Batenkov's established works but also to Iliushin's own poems. The task is, thus, reformulated as a *classification* problem.

4.2.1 Features

A full-scale replication of Shapir's study cannot be undertaken with large corpora because of the limitations of automated text analysis and scansion. We therefore confine our analysis to three levels:
— *Vocabulary* modelled by *lemmata frequencies* (with lemmatisation provided by MyStem 3.1, https://yandex.ru/dev/mystem/);
— *Morphology* modelled indirectly by character 3-grams (excluding punctuation and including blank spaces);
— *Versification* modelled by the rhyme features described in Section 2.2 (rhyme recognition provided by RhymeTagger (Plecháč 2018); IPA transcription provided by Espeak, http://espeak.sourceforge.net/). We do not consider rhythmic features because of the scarcity of lines in any particular metre in the data for either Batenkov or pseudo-Batenkov.

4.2.2 Fine-Tuning

Our first goal is to determine the most efficient feature space. To do this, we train multiple models with the following sets:
(1) frequencies of the *n* most common lemmata (L),
(2) frequencies of the *n* most common character 3-grams (G),
(3) frequencies of the *n* most common lemmata and the *n* most common character 3-grams (LG) and
(4) frequencies of the *n* most common lemmata and the *n* most common character 3-grams enriched with rhyme characteristics (LGR).

This is done for 40 different values of the most common types: $n \in \{50, 100, 150, \ldots, 2000\}$.

Here we use a corpus of Russian poems whose composition dates to the 1820s. This is partitioned into 200-line samples. Multiple poems can be combined in a single sample, and no poem contributes to more than one sample. This generates:

— 19 samples by Yevgeny Baratynsky,
— 23 samples by Mikhail Lermontov,
— 60 samples by Alexander Pushkin,
— 12 samples by Pyotr Vyazemsky,
— 36 samples by Nikolay Yazykov and
— 11 samples by Vasily Zhukovsky.

We apply the two different classifiers that will be used in subsequent experiments: linear SVM and cosine Delta.

To train the models, we follow the design laid out in Section 3.2. with five randomly selected authors and 10 randomly selected samples. Over 30 iterations, we perform cross-validation for the SVM model and nearest neighbour classification with the Delta approach.

The results can be seen in FIG. 4.4. The performance is similar to those recorded for other languages (Chapter 3): for all of the feature sets, accuracy generally increases to approximately the level of the 1000 most common types. At that point, it stabilises. For both classifiers, the LG combination tends to significantly outperform both L and G on their own. Even greater accuracy is almost always achieved, however, when rhyme features are also taken into account (LGR).

In the next set of experiments, we therefore retain LGR-based models and choose the 1000 most common types as the optimal level.

4.2.3 The One-Class Problem (Authorship Verification)

So far all the tasks we have considered in this book have involved authorship *classification*. In this situation, there is a closed set of candidates $\{A_1, A_2, A_3, \ldots, A_n\}$ and the goal is to determine which one is most likely the author of the text(s) X. In contrast, authorship *verification* deals with a different scenario. Here it is not possible to determine a closed set that we are sure includes the real author. The goal is instead to decide whether a certain A *is* or *is not* the author of X.

The Batenkov case needs to be treated first and foremost as a verification problem. If there are doubts about the origin of the disputed texts, then we first need to

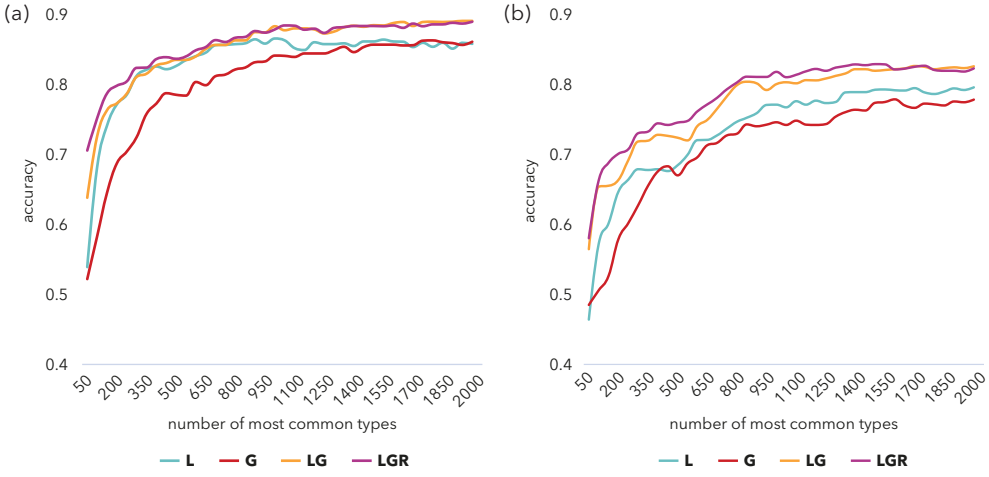


FIG. 4.4: Accuracy of (a) the SVM model and (b) the cosine Delta model with the most common lemmata (L), the most common character 3-grams (G), the L and G combination (LG) and the L and G combination enriched by rhyme features (LGR) across different levels of the most common types.

determine how likely it is that Batenkov himself wrote them regardless of Iliushin's status as a potential author. Here we loosely apply the unmasking technique (Koppel and Schler 2004; Koppel et al. 2007). In its classic version, this technique makes a series of pairwise SVM classifications between same-author and other-author samples. It then iteratively drops the most distinctive features from the learning process. Compared to other verification techniques such as those based on entropy or deep learning (Halvani et al. 2019), unmasking stands out for its clear assumptions and production of interpretable results.

Unmasking assumes that text samples from the same author will share deeper similarities than the samples of two different authors. In the former case, there may still be differences but they will emerge from high-level features such as theme, chronology or genre and not from the underlying style. Moreover, such features will inevitably be exploited by machine classification. That is why the original unmasking method relies on several stages of classification: in each iteration, a certain number of the most distinctive features are dropped and the classification is performed again. Given their underlying similarity, same-author samples should, thus, quickly become indistinguishable from one another while other-author samples retain their differences across many iterations. This is because their "distinctiveness" is distributed over many features and not concentrated in a few high-performing ones.

Since multiple poems can be combined in a single sample and no poem contributes to more than one sample, there is no reason to suppose that any high-level features distinguish the works of a single author. We therefore tweak the classic unmasking process by asking a simple question: Can the known Batenkov poems be distinguished from the disputed poems in a pairwise SVM classification?

To gauge the accuracy of this technique, we also test it on a control group of works published by other Russian poets in the 1840s and 1850s (i.e. the period when the majority of the disputed texts had allegedly been written). Like the Batenkov poems and the disputed poems, these works are divided into 100-line samples. (A 200 line size would generate only three samples from both Batenkov's work and the disputed poems). This produces:

— 13 samples by Mikhail Lermontov,
— 14 samples by Fyodor Tyutchev,
— 18 samples by Pyotr Vyazemsky,
— 15 samples by Nikolay Yazykov,
— six samples by Gavriil Batenkov and
— six samples from the disputed poems.

We then follow the four steps below:

(1) Randomly select 12 samples from each of the four "control" authors.
(2) Randomly split each group of 12 samples in half. These two groups are the A-samples and B-samples.
(3) Use the A-samples and the LGR feature set to train SVM models for each possible pair of "control" authors (i.e. Lermontov vs. Tyutchev, Lermontov vs. Vyazemsky, through to Vyazemsky vs. Yazykov). Perform *leave-one-out* cross-validation of each model.
(4) Train the SVM models with the LGR feature set for each "control" author using his own A-samples and B-samples as separate classes (i.e. Lermontov (A) vs. Lermontov (B), Tyutchev (A) vs. Tyutchev (B), Vyazemsky (A) vs. Vyazemsky (B), Yazykov (A) vs. Yazykov(B)). Perform *leave-one-out* cross-validation of each model.

We repeat this entire process 30 times for each quantity of the most common types: $n \in \{50, 100, 150, \ldots, 1000\}$. A new set of randomly selected samples is used in each iteration. For each n, we therefore obtain $4 \times 30 = 120$ accuracy estimations for samples written by the same author and $\binom{4}{2} \times 30 = 180$ accuracy estimations for samples written by different authors. Finally, for each n, we also cross-validate the Batenkov poems against the disputed poems model.

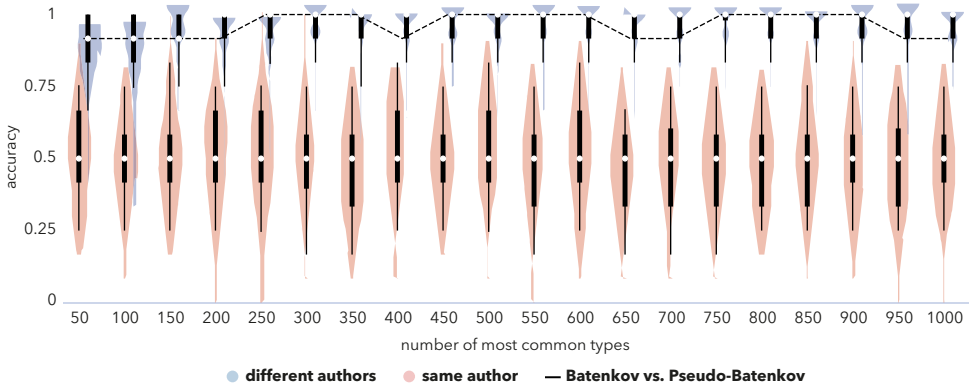


FIG. 4.5: Accuracy of pairwise classifications for different quantities of the most common feature types. Boxplots depict the median, the interquartile range (box) and the 5th-to-95th percentile range (whiskers).

FIG. 4.5 shows the results. The "control" authors behave as might be expected. The median classification accuracy for same-author pairs (A-samples vs. B-samples) hovers around 50%, meaning that on average they are indistinguishable for a classifier. At the same time, accuracy remains high for the pairwise classification of different authors as well. The dashed line in FIG 4.5. represents the classification accuracy for Batenkov poems vs. disputed poems. Without exception, this line follows the general trend for texts from two different sources.

Although these results seem fairly convincing on their own, we wish to go one step further and interpret them in terms of probabilities. As there appears to be no significant divergence among different quantities of the most common words (except perhaps when using the lowest values to classify different authors), we merge all of the values to obtain accuracy estimations for: (1) same-author classifications, (2) different-author classifications and (3) Batenkov poems vs. disputed poems classifications. A Mann-Whitney test[24] shows that the probability of these outcomes if Batenkov *was not* the author of the disputed poems is 0.9265 ($U = 111$, $n_1 = 3600$, $n_2 = 20$). In contrast, if Batenkov *was* the author, the probability is less than 10^{-14} ($U = 60618$, $n_1 = 2400$, $n_2 = 20$).

24 As there are always 12 samples, there are only 12 possible outcomes of cross-validation. The variable in question is, thus, not continuous but discrete. We therefore opt for the non-parametric Mann-Whitney test over the perhaps more expected *t*-test.

4.2.4 The Two-Class Problem (Batenkov vs. Iliushin)

There is, however, a fly in the ointment. As we have observed, Batenkov's poems spanned the 1810s to the 1860s with a significant gap from 1825 to 1846 when he was in solitary confinement (see FIG. 4.6 for a more detailed depiction of this output). The disputed poems date almost entirely from the period after his imprisonment. We therefore cannot rule out a scenario also raised by Shapir: during Batenkov's confinement, there might have been a dramatic change in his writing style which would explain the irregularities in the disputed poems. To address this objection, the disputed poems have to be compared with Batenkov's later poems alone.

Unfortunately, there are not enough data to perform a pairwise SVM experiment with only the poems that Batenkov published after his release. We therefore need to switch to the less data-hungry Delta method. We depend here especially on the cosine variation, which has proven to be the most reliable technique with our "control" authors. The problem is, thus, reframed as a classification task.

To begin, we increase the sample size to 200 lines. This produces the following numbers of samples per author:

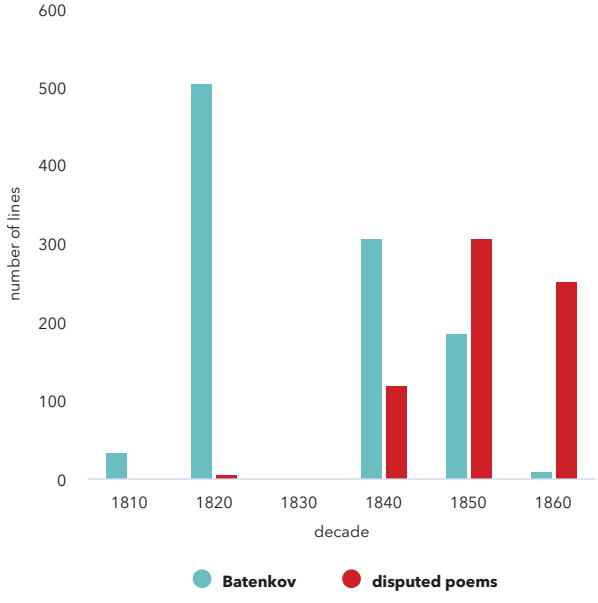


FIG. 4.6: Batenkov's poems and the disputed poems according to their (supposed) composition dates.

— Mikhail Lermontov (8),
— Fyodor Tyutchev (8),
— Pyotr Vyazemsky (12),
— Nikolay Yazykov (8),
— Gavriil Batenkov (2) and
— disputed poems (3).

Over multiple experiments with different feature space settings, the disputed poems remain clustered with Batenkov's poems. This does not say much about the Iliushin hypothesis, however, since the suspected author is not included in the candidate set (if, on the other hand, the disputed poems and Batenkov's poems did not cluster together, this might be interpreted as strong evidence of a forgery).

Although Iliushin never published any poems under his own name, preferring to mask his authorship of non-academic works, several texts have been attributed to him by consensus. These include *Дедушка и девушка* (published as an anonymous poem), *Michele Trivolis — Максим Грек* and *Добрый вампир* (both published under the name Y. F. Sidorin) and *Тайная дочь декабриста Бесстужева…* (the so-called Pseudo-Grigoriev, which was presented as a work by the poet A. Grigoriev, 1822–1864). All of these are long narrative poems from which it is possible to extract a sample comparable to those used in our past experiments.

Now we add the (apparent) Iliushin samples to the corpus and perform another battery of experiments. The quantity of most common types is set to 1000 for both lemmata and character 3-grams. To verify the robustness of these results, we perform 10,000 classifications; in each iteration, 0–1000 types of each feature are dropped from the classification (both the quantity of types and the features themselves are randomly selected). The results are summarised in a confusion matrix (TAB. 4.3).

	Batenkov	Iliushin	Lermontov	disputed poems	Tyutchev	Vyazemsky	Yazykov
Batenkov	**1**			0.06			
Iliushin		**0.99**	0.01	0.21		0.09	
Lermontov			**0.89**		0.03		
disputed poems			0.02	**0.73**			
Tyutchev			0.03		**0.95**	0.01	
Vyazemsky		0.01	0.04		0.01	**0.89**	
Yazykov			0.01		0.01		**1**

TAB. 4.3: Confusion matrix (relative counts). Rows represent the author predicted by the model while columns represent the actual author. Individual cells show the relative number of predictions in each case.

In over 20% of the vector spaces, one sample of the disputed poems appears to be closer to Iliushin's poems than to the other disputed poem samples. This is completely unlike the pattern with the other authors, which showed only minimal variation across the predictions.

Interestingly enough, all of these "misattributions" of the disputed poems to Iliushin concern just two of his samples. These are both poems published under the name Y. F. Sidorin. This, in turn, raises a question: Do these works differ somehow from the other two Iliushin samples?

There are indeed several differences beginning at the level of metre. The Sidorinian poems are written in iambic pentameter, one of the most common metres in Russian poetry in the first half of the 19th century; in contrast, *Дедушка и девушка* is loosely trochaic and "PseudoGrigoriev" is dactylic. Clearly, vocabulary, morphology and rhyme structure can all be profoundly affected by the choice of metre as well.

A closer look at the Sidorinian poems yields even more information. FIG. 4.7 shows the cosine distances across various quantities (50, 100, 150, …, 2000) of the most common types when the disputed poems are compared with (i) the Sidorinian poems and (ii) Batenkov's own poems published between the 1840s and the 1860s. In all of the

FIG. 4.7: Cosine distances between the disputed poems and (1) the Sidorinian poems (Iliushin) and (2) Batenkov's own poems published between the 1840s and the 1860s for different quantities of the most common types.

vector spaces defined by up to the 1000 most common types, the Sidorinian poems appear to be closer to the disputed poems than Batenkov's own texts are. Then, after a spell in which the distances are more or less even, Batenkov becomes the preferred candidate. This, in fact, seems to be precisely the behaviour we would expect from a forger. Wouldn't such a person imitate an author's obvious idiosyncracies (low-frequency features) but fail to adopt the less obvious ones (high-frequency features like function words and common suffixes)?

4.2.5 Summary

These results do not leave much scope for agreement with Shapir about the essential unverifiability of the disputed poems. As we have seen, when we attempt to treat these texts like original works by Batenkov, they behave radically differently from what we would expect of 19th-century poems written by a single author. Moreover when we try to classify them, they are mistaken for Iliushin's original poetry far more often than they are for the works of their alleged author.

These findings, however, should not be treated as definitive proof of a forgery. After all, stylometry never delivers definitive answers. We are always left with some uncertainty about classification accuracy. The disputed texts may include some unknown original lines later heavily edited or rewritten by the custodians of Batenkov's manuscripts or those who came to study them. And indeed Batenkov may have survived some personality-altering experiences that suddenly rewired his writing habits. Since, however, we have found no evidence to support these possibilities, we would suggest that from now on the null hypothesis should be that "the Batenkov and pseudo-Batenkov texts were not written by the same author".

In practical terms, our results are not surprising since so many scholars and readers remain convinced of Iliushin's forgery despite Shapir's insistence on indeterminate authorship. There are, however, larger theoretical questions at stake: Does language reflect an author's identity? Can a reader recognise the distinctive features of literary style? Are these stylistic features associated with authorship?

Shapir (2000) writes: "Anything conceived by chance, which is unique and unrepeatable, cannot be compared; anything stable and recurring can be abstracted and replicated" (419). The whole history of stylometry reflects an ongoing quest for a means to compare the unique. The methods we rely on seek to access low-level linguistic features that vary greatly among individuals, who usually do not exercise conscious control over them.

Still, stylometry does not give us access to literary forms or any perceived abstract features of a text. How much can we know about Batenkov's literary techniques by observing the distinctive linguistic features of his poetry? Perhaps something from a handful of nouns and verbs, less from pronouns and adverbs and next to nothing from his habit of ending rhymes with a particular sound [x] and his overuse of the character bigrams "ви" and "ен".

Shapir's words speak to the hope of finding the author's identity in linguistic phenomena that can be conceptualised and connected back to literary forms. His work on Batenkov reportedly failed to show this: the authorial signal became fuzzy and the results remained inconclusive. Shapir's uncertainty may find support from recent studies that show that differences in how literary contemporaries use cultural forms and devices ("anything that can be replicated") may be negligible and incomparable to the gigantic gaps between them in the literary market or academic canon (Moretti 2013: 145–147; Porter 2018; Sobchuk 2018: 91–97). Stylistic identity is not, however, bound to any skewed power-law distribution: unlike fame, critical attention and other goods of the symbolic economy, it is distributed equally across the population. That is why stylometry works: everyone who writes is an author.