

Rolling stylometry

Maciej Eder

Polish Academy of Sciences, Institute of Polish Language, and
Pedagogical University of Kraków, Kraków, Poland

Abstract

This article introduces a new stylometric method that combines supervised machine-learning classification with the idea of sequential analysis. Unlike standard procedures, aimed at assessing style differentiation between discrete text samples, the new method, supported with compact visualization, tries to look inside a text represented as a set of linearly sliced chunks, in order to test their stylistic consistency. Three flavors of the method have been introduced: (1) Rolling SVM, relying on the support vector machines (SVM) classifier, (2) Rolling NSC, based on the nearest shrunken centroids method, and (3) Rolling Delta, using the classic Burrowsian measure of similarity. The technique is primarily intended to assess mixed authorship; however, it can be also used as a magnifying glass to inspect works with unclear stylometric signal. To demonstrate its applicability, three different examples of collaborative work have been briefly discussed: (1) the 13th-century French allegorical poem *Roman de la Rose*, (2) a 15th-century translation of the Bible into Polish known as *Queen Sophia's Bible*, and (3) *The Inheritors*, a novel collaboratively written by Joseph Conrad and Ford Madox Ford in 1901.

Correspondence:

Maciej Eder, Polish Academy
of Sciences, Institute of
Polish Language, al.
Mickiewicza 31, 31–120
Kraków, Poland.

E-mail:

maciejeder@gmail.com

1 Introduction


In classical approaches to stylometry—be it authorship attribution, genre recognition, or, say, a distant-reading classification of hundreds of novels—the goal is to compute a measure of similarity between the texts in a corpus, in order to discover hidden patterns or regularities. In authorship attribution, it involves extracting the authorial profile from a disputed text, followed by a procedure of identifying the best match in a set of ‘candidates’; in stylometry beyond attribution, it is aimed at finding groups of stylistically similar works. Even if the input texts are split into samples, the basic high-level stylometric unit is a literary work in its entirety. The hypothesis that consecutive sections of a given text might reveal linear development of

certain stylistic features is still a relatively new perspective in this field.

Sequential analysis is a very attractive way of assessing linear phenomena; it is widely used in signal processing, econometrics, weather forecasting, electroencephalography, and so forth. It relies on the general assumption that the sequential order of elements is as important as the elements themselves: if a given series of events is nonrandom, the next element in the series should be—to some extent—modelable and hence predictable. Since natural languages are linear by definition, it was only a matter of time before the methods of sequential analysis were adopted to linguistics. In a study on language as a probabilistic phenomenon, Herdan introduces a classical distinction between two domains of quantitative linguistics: ‘language in the mass’ versus

‘language in the line’ (Herdan, 1966, p. 423). While the former category refers to the popular ‘bag-of-words’ approach, the latter emphasizes the importance of—among other things—the preceding context of analyzed words.

Interestingly, the intuition that language is a modelable sequence of nonrandom units was verbalized as early as in 1913, in the fundamental study introducing Markov chains: to test mathematical assumptions of his new method, Markov used sequences of letters from *Eugene Onegin* by Alexander Pushkin (Petruszewycz, 1981; Markov, 2006 [1913]). Even if this particular contribution to linguistics was symbolic rather than significant, Markov’s studies had a great impact on theoretical foundations of sequential methods. In particular, the concept of the moving (sliding) window needs to be mentioned here.



This relatively simple idea revolutionized the way in which linear phenomena could be assessed: supposing that a sequence of events (referred to as time series) consists of N elements, the goal is to measure mathematical properties of a subset of k consecutive elements extracted from the beginning of the sequence, and then to move such a ‘window’ of the size k through the entire time series until the position $x = N - k$ is reached. In consequence, one obtains insight into particular segments of the data set in their development. It allows us to detect periodic regularities in the time series on the one hand, and possible disturbances or local idiosyncrasies on the other. The idea of moving window will be the key concept in the present article.

Sequential methods, including time series analysis, spectral analysis, and Markov models, have been widely used in natural language processing, including part-of-speech (POS) tagging, parsing, speech recognition, and synthesis, to name but a few applications. In the field of stylometry, autoregression models proved effective to assess versification (Pawłowski, 1999; Pawłowski and Eder, 2001). Markov chains—at character level—have been introduced to authorship attribution (Khmelev and Tweedie, 2001). Some elements of sequential analysis have been implicitly used in several attribution studies involving word n -grams, character

n -grams, or POS n -grams as discriminative features (Hirst and Feiguina, 2007; Stamatatos, 2009; Koppel et al., 2009; Eder, 2011), since n -grams are in fact sequentially ordered series of nonrandom units.

The concept of moving window has been extended by van Dalen-Oskam and van Zundert in their study on the medieval Dutch Arthurian epic poem entitled *Roman van Walewein* (van Dalen-Oskam and van Zundert, 2007). Instead of analyzing sequences of three or so letters, the authors used a very high-level moving window of the length of several hundred words. The aim of such an approach was to generate a series of virtual subsamples from the *Walewein* in order to test their stylometric consistency throughout the whole text. Other notable approaches to visualize stylistic shifts using moving windows include a paper on Middle-Dutch rhyme words (Kestemont, 2010), on three disputed English prose texts (Burrows, 2010), and on *The Tutor’s Story* by Kingsley and Mallet approached with t -tests (Hoover, 2011).

Similar studies have now been made possible by the recently introduced Rolling Delta method, available in the R package ‘Stylo’ (Eder et al., 2013). The technique has been applied to examine collaborative works by Joseph Conrad and Ford Madox Ford (Rybicki et al., 2014), and used in a benchmark study on Dickens (Tabata, 2014). In this method, the standard windowing procedure is run throughout a reference corpus: a representative centroid for each reference text that consists of the mean relative frequency for each of the N words in the windows extracted from the text is calculated. Next, the test text is also divided into windows and a distance measure (in this case, Burrows’s Delta) between each text window and each reference centroid is computed. The results are visualized using a set of curves—one for each reference text. The final step involves identifying, for each window, the lowest line, i.e. the most similar reference text. Whenever a takeover (line crossing) occurs, the respective window is assumed to reveal a stylistic change.

Regardless of the level of mathematical complication of the aforementioned sequential stylometric techniques, the basic underlying concept they share is quite simple. Namely, the goal is to split an input text into several subsequent samples (windows) and

to contrast them one by one against the reference corpus. It is crucial, however, to keep the original order of the analyzed samples.

Arguably, a generalization of the above idea is straightforward. Since the windowing procedure can be used as a framework for simple similarity measures, such as *t*-tests, it will be also, by extension, applicable to any machine-learning classifier. The present study is aimed at discussing such a generalization and at introducing a robust method of rolling classification. The method will use the concept of moving window in combination with standard supervised classification techniques.

The article is divided into two parts. In the first part, theoretical assumptions of the new method are discussed; they concern sampling issues (moving window), classification (three different machine-learning techniques), and visualization. In the second part, exemplary applications of the rolling method will be presented. The chosen examples include (1) the 13th-century French allegorical poem *Roman de la Rose*, (2) a 15th-century translation of the Bible into Polish known as *Queen Sophia's Bible*, and (3) *The Inheritors*, a novel collaboratively written by Joseph Conrad and Ford Madox Ford in 1901.

Since the goal of the article is to introduce a new technique rather than to perform an extensive benchmark, the presented applications are not case studies in a strict sense. In the first case, there is no reference corpus available, the second example is a highly collaborative work at multiple levels—probably too complex to be reliably solved—while the third application's results substantially depend on the input parameters. Obviously this is not a good material for a systematic benchmark. The chosen examples, however, aim at showing the research questions that can be assessed using the new technique; they are also used to point out a few methodological issues that need to be solved in the future.

2 Method

2.1 Sampling

Supposing there is a work written collaboratively—i.e. a text in which some authorial takeovers are

suspected to have happened—the procedure should start with chunking the text into consecutive samples. It is true that any already-existing sections (chapters, acts, cantos, and so forth) can be used as natural chunk delimiters: this solution is obviously preferable when external evidence suggests such a character of authorial collaboration. However, sequential methods show their real power when the samples are distributed evenly throughout the input data set. In such a case, the data set shows the properties of a time series, and thus it can be analyzed using dedicated tools, e.g. one can estimate the autocorrelation function to see if there are any cyclic regularities. Also, being a representation of actual timeline, the internal development of textual units can be easily visualized and reliably compared. For this reason, it is better to split the text in question into equal-size blocks of *N* words (tokens). The desired sample size is a key parameter here, and it needs to be decided arbitrarily. On the one hand, keeping this parameter small increases the resolution of sampling—an essential factor to pinpoint stylistic breaks—on the other hand, however, below a certain sample size, classification methods become blind. It has been shown that minimal sample length for authorship attribution is roughly 5,000 words, depending on the language and the genre examined (Eder, 2013).

The resolution of sampling can be substantially improved when the concept of moving window is applied. Its inherent feature is that it allows sample overlapping, so that some observations in a data set can be reused several times. In the case of rolling stylometry, it means that a sequence of tokens from the end of sample *A* will reappear in the middle of sample *B*, and at the beginning of sample *C*. The idea of overlapping is shown in Fig. 1, where different types of sampling—with and without overlap—are represented. The only difference is that in classical approaches, e.g. in Markov models, the length of the moving window rarely exceeds 2–3 units (letters or syllables), and consequently the sample overlap is also very small. Stylometric windows, on the other hand, rely on extremely wide windows of hundreds of words, and thus their overlap needs to be augmented accordingly. In the exemplary applications discussed below, the

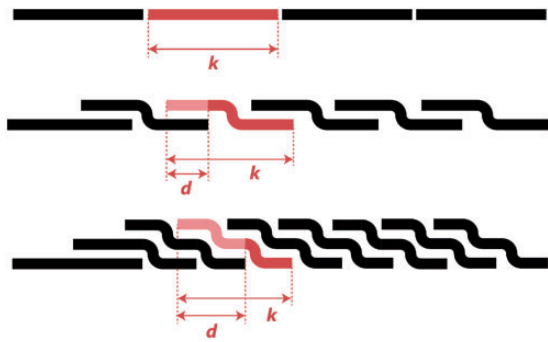


Fig. 1 Chunking a text into subsequent samples using a windowing procedure: in the above three variants, the same sample size of k elements is used in combination with different sample overlap (denoted by d)

window size of 5,000 words has been used, with an overlap of 4,500 words. Arguably, other combinations of the window size and the overlap parameters will lead to (slightly) different final results. This issue needs to be examined systematically, not only taking into account chunking options, but also a few style-markers (POS-tags, letter n -grams, and so forth), feature normalization algorithms, as well as alternative classifiers. Such a controlled benchmark experiment exceeds substantially the scope of this article.

2.2 Classification

The observation that the chunks produced by the windowing procedure can serve as regular text samples for authorship attribution is quite obvious. In short: instead of attributing a given text in its entirety, the goal is to perform an independent similarity test for each chunk, and then to inspect the results as a sequence of ordered stylistic signals.

Arguably, any classification method can be combined with the above framework. In the present study, however, three supervised classification techniques known for their high accuracy will be used. These are support vector machines (SVM), nearest shrunken centroids (NSC), and Delta in its classical Burrowsian flavor. These classifiers have been thoroughly tested in authorship attribution (Burrows, 2002a; Hoover, 2004; Koppel et al., 2009; Jockers and Witten, 2010; etc.).

Even if they rely on substantially different mathematical kernels, SVM, NSC, and Delta use exactly the same corpus setup to carry out the classification. Namely, a number of representative samples for each class is expected to constitute a reference set (training set), while the remaining samples, including anonymous ones, go to a test set. Next, each sample from the test set is checked against the training set in order to identify the most similar authorial profile (i.e. the best matched training class). Unlike SVM and NSC, however, Delta does not combine individual training samples into averaged profiles for each class. This means that for a two-class problem Delta produces a ranking of the most similar training samples, beginning from the best match, e.g. $\{A_c, A_a, B_x, A_b, B_z, B_y, \dots\}$, while a standard classifier provides a ranking of composite classes: $\{A_{abc}, B_{xyz}\}$ or $\{B_{xyz}, A_{abc}\}$ (i.e. the most probable class followed by the less probable one). This feature of Delta will be used below (see Section 3.3) to visualize the first, the second, and the third hint of the classifier at a time, even for two-class problems. However, such a peculiarity does not affect the general setup of the experiment, which is shared across the methods. In rolling stylometry, the same general setup is used as well. The only difference is that the test set contains a single work to be chunked automatically into equal-sized segments.

Worth mentioning is the fact that multidimensional methods are very sensitive to the number of features used. In stylometry, the commonly accepted type of features (style-markers) is frequencies of the most frequent words (MFWs), even if there is no consensus how many MFWs should be used. The rolling techniques discussed in this article can be used either with MFWs, or with alternative style-markers, such as character n -grams, POS tags, or even manually selected content words. The number of features can be customized too. To keep things simple, the experiments discussed below were carried out using MFWs, with three different frequency ranges of 100, 500, and 1000 MFW (one range at a time).

2.3 Visualization

The final stage of the analysis involves a graphical representation of stylistic changes throughout a

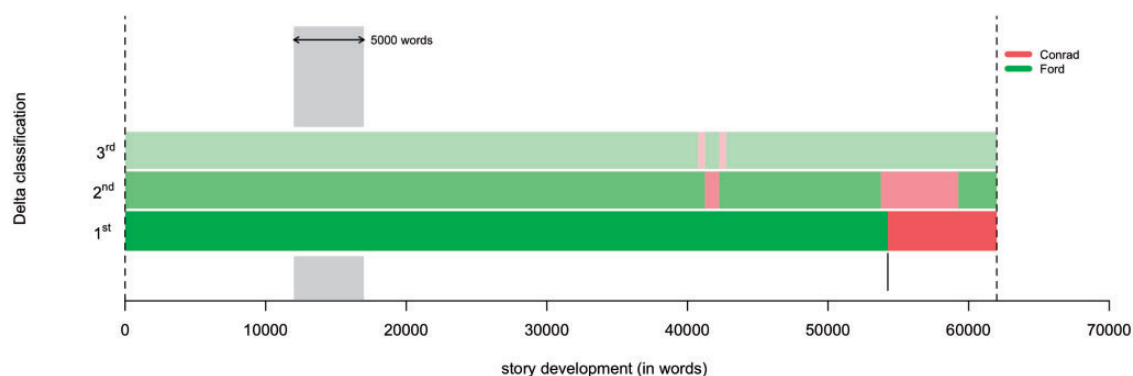


Fig. 2 ‘The Inheritors’ by Conrad/Ford assessed using Rolling Delta and 1,000 MFWs. The bottom stripe indicates the first ranked candidate (i.e. the most probable), then comes the second and the third suggested class

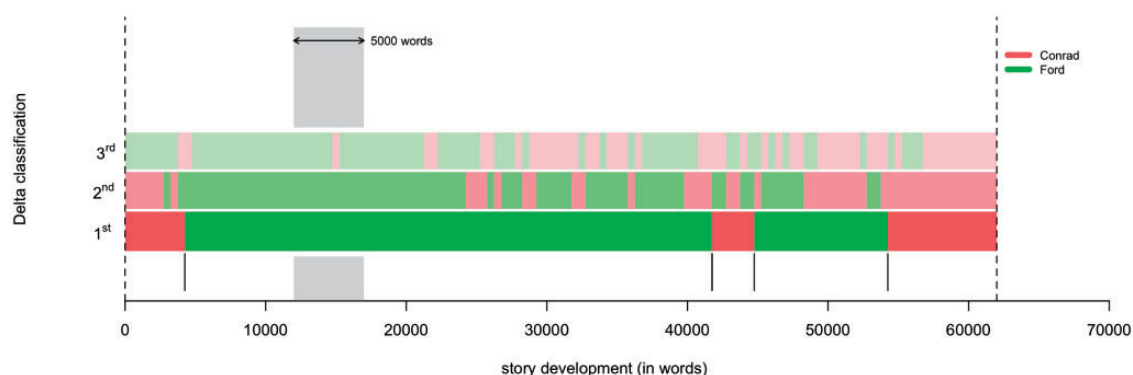


Fig. 3 ‘The Inheritors’ by Conrad/Ford assessed using Rolling Delta and 500 MFWs

chunked text. It is true that standard classification techniques do not need any visualization (they provide a list of assigned classes, which is informative enough *per se*). However, one cannot deny an escapable explanatory power of graphs, trees, and diagrams. For that reason, a simple home-brew graph has been added.

The design was based on the assumption that simplicity improves visual informativeness. Thus, to keep the plot clean, any redundant information has been removed. The goal was to emphasize visually the most likely candidate—i.e. the actual answer of the classifier—and to keep less probable candidates slightly in the shadow. To this end, horizontal stripes colored according to the assigned classes were used, the primary stripe bolded. In pure

form, this idea is implemented in the Delta variant of the method (Figs. 2–4). For each chunk, the first, the second, and the third ranked candidates are represented by an appropriate (tiny) segment of the bottom stripe, the middle stripe, and the top one, respectively. When classification results are consistent across a number of chunks, the stripes tend to be unicolored rather than patchwork-like.

Simple as it is, the visualization does not provide a good cue as to how reliable the output is—this is due to inherent limitation of the Delta method. In the case of SVM and NSC, however, the output has been slightly enhanced: it seemed reasonable to take advantage of the final probability scores these methods optionally provide. In standard SVM, class assignment is made according to decision values,

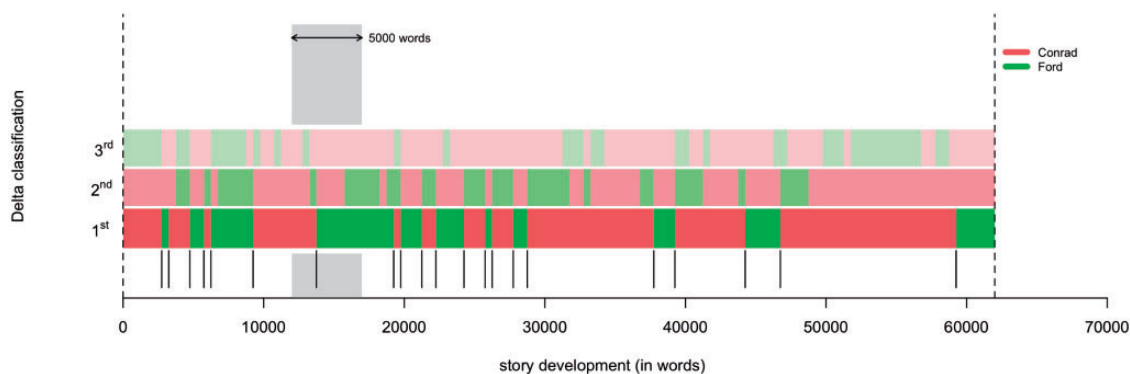


Fig. 4 ‘The Inheritors’ by Conrad/Ford assessed using Rolling Delta and 100 MFWs

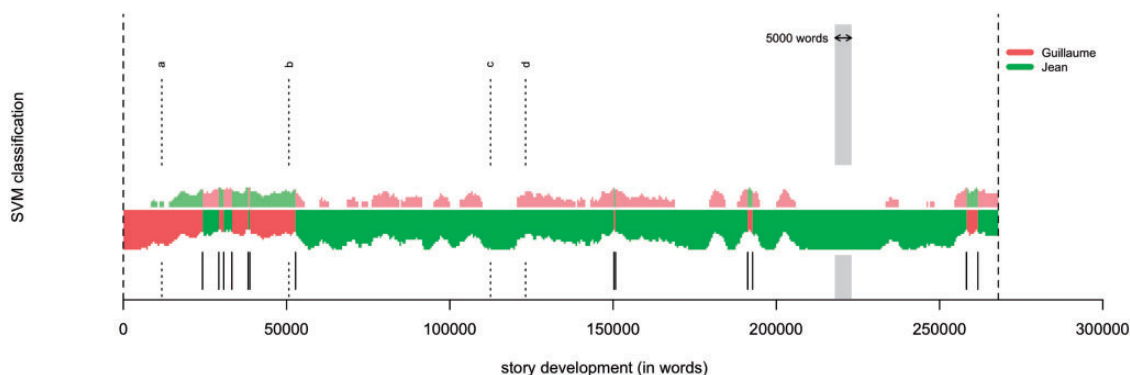


Fig. 5 ‘Roman de la Rose’ assessed using Rolling SVM and 100 MFWs. The level of certainty of the classification is indicated by the thickness of the bottom stripe. The commonly accepted division into two parts of the poem is marked with the vertical dashed line ‘b’

which can be either negative or positive: the higher they are, the more robust the assignment is. For instance, if the decision values for a given sample are as high as 0.93 in favor of the class A, and -0.93 of the class B, the sample is attributed to A with a high degree of certainty. If the values are 0.01 and -0.01 , respectively, then the sample is still attributed to A, but such a classification is not very robust. Now, if one normalizes the decision values, they can be used to control the width of the plotted stripes (Figs 5 and 6). In consequence, a stripe wide in its bottom part stands for a robust classification, and the other way around: the more a given segment is overtaken by the secondary (i.e. top) stripe, the

more dubious the method. Since the normalized values always add up to 1, the adjusted widths of the two stripes make them look like a single band, gently waving up and down the reference line.

In NSC, regular final probabilities were used. Adopting them to control the width of the stripes turned out to be straightforward (Figs 7 and 8). At first glance, the stripes produced by NSC and SVM are significantly different: the former flow gently along the x axis, the latter are more solid, with some ragged areas. This is because NSC is generally less hesitant in classification than SVM. One might say, NSC is devoted to the phrase ‘let your statement be, “Yes, yes” or “No, no”’; anything beyond these is

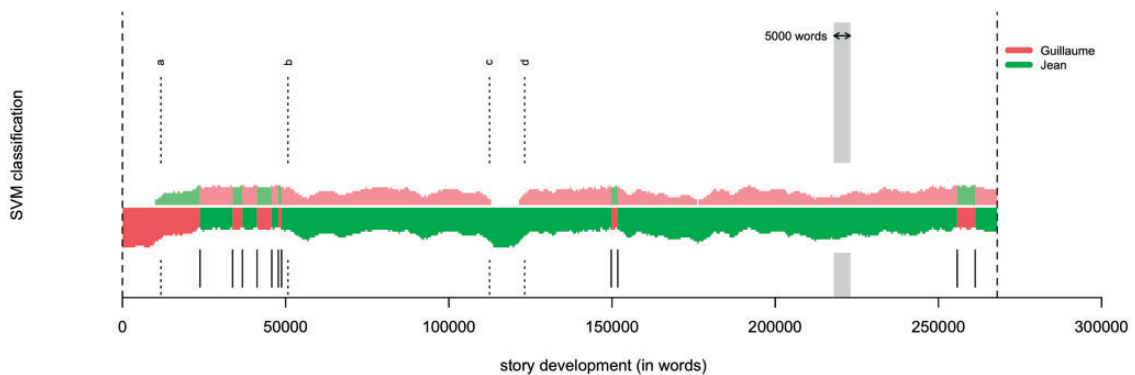


Fig. 6 ‘Roman de la Rose’ assessed using Rolling SVM and 500 MFWS

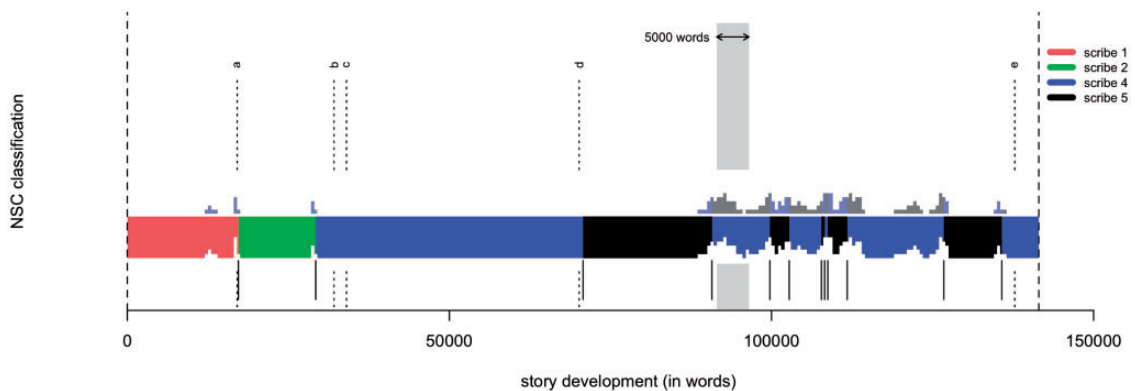


Fig. 7 ‘Queen Sophia’s Bible’ assessed using Rolling NSC and 100 MFWS

of evil’ (Matthew 5.37). However, even if sometimes too self-confident, NSC is still one of the best classifiers for stylometry (Jockers and Witten, 2010).

To show the above variants of the rolling technique in action, three exemplary applications are discussed in the following sections. For each case, a different flavor of the method is used: Rolling SVM, Rolling NSC, and Rolling Delta, respectively.

3. Exemplary Applications

3.1 Roman de la Rose

This 13th-century French poem, styled as an allegorical dream vision, is a perfect material to test sequential stylometric methods. There is a

consensus among scholars that the first part of the poem has been written by Guillaume de Lorris around 1230, and the second part has been completed by Jean de Meun about the year 1275. More importantly, unlike the Middle-Dutch poem *Roman van Walewein*, which has also been written by two authors in the 13th century, the takeover point in *Roman de la Rose* is well known: Guillaume de Lorris is the author of the opening 4,058 lines (ca. 50,000 words), and the second part by Jean de Meun consists of 17,724 lines (ca. 218,000 words). This knowledge is supported by the text itself. Namely, Jean de Meun explicitly writes about the collaborative authorship of the poem, he indicates the name of his predecessor as well as his own name, and he points out the last lines written by Guillaume de Lorris.

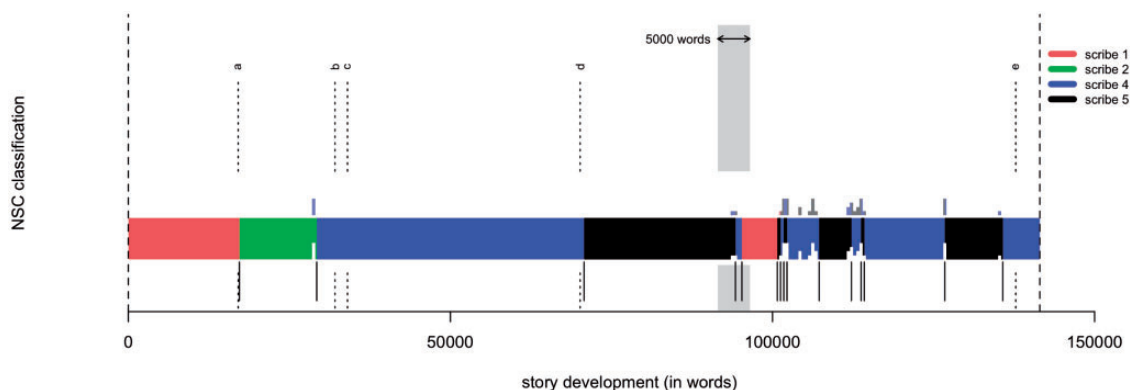


Fig. 8 'Queen Sophia's Bible' assessed using Rolling NSC and 500 MFWs

The research hypothesis is quite simple in this case: a given classification method is to be called effective if it captures the authorial takeover. Also, one can expect that an effective method reveals a clean segment for the first author followed by an equally clean segment for the second one. Any contamination within the two parts should be considered suspicious.

There is a fly in the ointment, though. Straightforward at first glance, this case contains a nontrivial issue, since there are no extant texts written by the first author of *Roman de la Rose* that could serve as a comparison corpus. In their seminal study on *Walewein*, the authors faced exactly the same problem (van Dalen and van Zundert, 2007). The proposed solution was rather tricky: two fragments from *Walewein* itself were copied to form the reference corpus. One sample was transplanted from the beginning of the poem, the other from its final part.

The same method—with the same caveat—can be used to assess *Roman de la Rose*: 10,000 words (roughly 1,000 lines) from the beginning, and the same amount of data from the middle (words from the range of 113,000–123,000) have been copied to serve as a surrogate for the training corpus. To carry on the analysis, the edition by Marteau, slightly outdated but open-accessibly available at the Gutenberg Project Web site, has been used (Marteau, 1878). The results obtained using Rolling SVM and 100 MFWs are shown in Figs 5 and 6. The first observation is that the takeover—marked with the

dashed vertical line 'b'—has been more or less precisely recognized. Also, a vast majority of the chunks have been robustly attributed to their actual authors; certainly, the sections that were transplanted to the training corpus, delimited with the vertical lines 'a' and 'c–d', are correctly recognized as well. In the remaining sections, the attribution accuracy depends on the analyzed author: the part by Jean de Meun reveals a more consistent authorial signal, while Guillaume de Lorris's chunks are significantly cluttered in the middle of his share.

There are at least three possible explanations of these authorial inconsistencies. Firstly (and pessimistically), the method is not precise enough. Secondly, the transplanted sample from the beginning of the poem did not contain enough information about the authorial profile of Guillaume de Lorris. Last but not least—*horribile dictu*—the second author did some corrections in the Guillaume's passages. Wrongly attributed chunks at the end of the poem are also interesting: they seem to show that Jean de Meun put less stylistic effort (in the sense that the authorial signal is weak) when his work was about to be finished. It should be emphasized here that very similar local misclassifications were noticed using the remaining methods: Rolling NSC and Rolling Delta.

3.2 Queen Sophia's Bible

The first known translation of the Bible into Polish was completed in the year 1455. It was commissioned by Sophia of Halshany, wife of Władysław

II Jagiełło, King of Poland, and Grand Duke of Lithuania. The translation was not intended for liturgical purposes, neither was it known outside the royal court. The text was handwritten on parchment in two large codices. The second volume has been disintegrated quite early to provide parchment for book covers (a few folios have survived), and the first volume, containing a good share of the Old Testament, was lost during the World War II, probably damaged. Luckily enough, however, a black-and-white facsimile of the extant manuscript was published in 1930.

Queen Sophia's Bible is a very interesting example of a collaborative (at many levels) work. It is translated directly from Czech rather than from the Vulgate, but it is difficult to decide which of the several variants of the Czech Bible was used (Deptuchowa, 2008, pp. 9–12; Wanicowa, 2009, pp. 76–81). Textual similarities suggest that the beginning passages were translated from one of the Bibles of the oldest 1st redaction, while the next passages were probably rendered after the 2nd redaction—it should be emphasized, however, that despite differences, all the Czech redactions are ultimately derived from the same translation (Kyas, 1997). When it comes to textual source, then, *Queen Sophia's Bible* is a translatorial composite: the original Hebrew/Greek Bible was translated into Latin (the Vulgate), then into Czech, then it was modernized (2nd Czech redaction), and then rendered in Polish.

Closer examination of the codex (or its extant facsimile, to be precise) reveals another level of collaboration: five different scribal hands can be distinguished quite easily. The scribal takeovers can be also noticed at the level of orthography. Moreover, the sections for particular scribes seem to be different at stylistic level as well: the quality of translation in the sections written by the 2nd and 5th scribal hands are considered to be smooth, while the 1st hand is claimed to be 'very literal' and generally of poor quality (Urbańczyk, 1933). Thus, it has been hypothesized that several translators were involved in rendering the Biblical text or, alternatively, that the scribes were at the same time translators.

All the above variables taken together make *Queen Sophia's Bible* a multifaceted collaborative

work, in which the translatorial, authorial, and scribal signals are heavily mixed. Certainly, telling these signals apart is unrealistic. What seems feasible, however, is corroborating the hypothesis that the scribal takeovers are correlated with stylistic (i.e. translatorial) transitions. From stylometric point of view, translators are known to be (almost) invisible. However, even if this is the author of the original that is usually stylistically predominant in a translated text, a few successful translatorial attributions show that such a presence-in-a-shadow can also be pinpointed (Burrows, 2002b; Rybicki, 2012; Rybicki and Heydel, 2013). In the case of *Queen Sophia's Bible*, the voice of the original Hebrew authors is covered by so many layers that the risk of its predominance is lower than in usual translations, and thus any stylistic changes in the Polish text might be, with a reasonable probability, attributed to the translator rather than to the original author. Two issues are to be resolved, though. Firstly, the scribal signal has to be neutralized in order not to interfere with the stylistic one. Secondly, a comparison corpus has to be compiled.

It is a truth commonly known that in medieval manuscripts, orthographic variants are innumerable and they highly depend on scribes' preferences, education, particular handwriting school, and so forth. Certainly, the same applies to the old-Polish language and to *Queen Sophia's Bible*. Being a strong discriminator, orthography can be used to tell the scribes apart (Thaisen, 2012), but at the same time it weakens—or interferes with—the stylistic signal. It has been shown, however, that the impact of scribes' voices can be neutralized by dealing with a corpus in modernized transcription rather than in transliteration (Kestemont and van Dalen, 2009). Since *Queen Sophia's Bible*, along with a dozen of other 15th-century Polish texts, has been recently edited in a form of a transliterated and transcribed parallel corpus (Twardzik, 2006), the suggested solution could be immediately applied.

When it comes to the second issue: a standard reference corpus cannot be compiled for the same reasons as in *Roman de la Rose*, but, similarly, the takeover points are known very well. Thus, it was possible to reliably transplant a few samples into the training corpus. Segments of 10,000 words (one

segment per class) have been transplanted from the parts by the 1st, 2nd, 4th, and 5th scribe, i.e. roughly 50%, 50%, 25%, and 10% of the material by these scribes, respectively. The 3rd scribe's contribution (mere 2,000 words in total) is too short to allow any sample extraction. Certainly, since one training class is missing, this short segment will be by definition misclassified.

The results for Rolling NSC are shown in Figs 7 and 8; handwriting changes are marked with vertical dashed lines. Regardless of the number of MFWs tested, the classifier detects a few style breaks in the data set. The most valuable result, however, is that the stylistic breaks take place in parallel with scribal takeovers (with some minute shifts). This is a strong evidence in favor of the 'many translators involved' hypothesis. Interestingly, the third stylistic break took place still in the 2nd scribe's section (i.e. between the markers 'a' and 'b'), as if a newly hired translator was forced to work with his predecessor's scribe for a while. Alternative explanations of this particular takeover are also possible, though: it might be a stylistic break in the original, e.g. the point where the 1st Czech redaction has been replaced with the new-fashioned 2nd redaction, it might be a break in the Latin pre-original, and so forth: one should remember that this case is stylometrically far too complex to make any conclusive statements.

Stylistic inconsistencies in the second half of the text should also be commented on. This part, written by the 5th scribe and recognized to be stylistically more or less consistent, contains some apparently misclassified segments. In Fig. 7, representing the results for 100 MFWs, some of the chunks seem to be partially eclipsed, and some are entirely overtaken by the 4th scribe's class. In Fig. 8, for 500 MFWs, an even less probable attribution to the 1st scribe appears, along with a lengthy section robustly yet wrongly attributed to the 4th scribe. This can be interpreted as a weaker stylistic voice in the chunks in question (or yet another translator involved), but at the same time it is a meaningful caveat: when the number of classes is limited and the open-set case cannot be ruled out, false positives might appear.

Close reading of the text itself provides a convincing explanation of the wrongly attributed final

chunks. Unlike the already-discussed misclassifications, which are rather accidental than systematic, the final chunks are robustly attributed to the 4th scribe regardless of the method used (Rolling SVM, Rolling NSC, Rolling Delta). The reason of the apparent blindness of the method turned out to be embarrassingly trivial. Namely, at the end of the critically edited text, all the extant fragments from the second volume of *Queen Sophia's Bible* are collected. As a result of corpus setup error, these concatenated fragments produced a fake stylistic signal, marked in Fig. 8 with the vertical dashed line 'e'.

3.3 Conrad versus Ford revisited

The last example, aimed at introducing Burrows's Delta as a rolling classifier, is a replication of the experiment conducted by Rybicki, Kestemont, and Hoover, reported in their study on Joseph Conrad's and Ford Madox Ford's collaboration (Rybicki et al., 2014). Having scrutinized three novels that were written collaboratively by Conrad and Ford, the authors of the paper conclude: 'The decisive domination of Ford's style over Conrad's in *The Inheritors* and *The Nature of a Crime* is interesting, as it seems to have survived Conrad's extensive editing that is confirmed by biographical evidence' (*ibid.*, p. 429). To test the above claim, the same reference corpus and the same test text have been used. Namely, *The Inheritors* has been compared with six novels by Conrad (*The Nigger of the Narcissus*, 1897; *Heart of Darkness*, 1898–99; *Lord Jim*, 1900; *Chance*, 1913; *Under Western Eyes*, 1911; *Victory*, 1915) and six novels by Ford (*The Benefactor*, 1905; *Privy Seal*, 1907; *An English Girl*, 1907; *Mr. Apollo*, 1908; *Ring for Nancy*, 1913, *The Good Soldier*, 1915).

The results for a very long vector of 1,000 MFWs confirm the findings of the previous study (Fig. 2), in which 1,000 MFWs were used as well. Apparently, it suggests that *The Inheritors* indeed was mostly written by Ford, with some final sections contributed by Conrad. However, when the MFW parameter is lowered, the picture of the collaboration changes substantially. For 500 MFWs, Conrad becomes quite visible, especially in the background (the second and the third suggestion of the classifier). When the number of MFWs is further reduced to 100 or so,

Conrad boldly comes out from the shadow to take a leading role in the duo's collaboration.

Very similar results were obtained using two other classifiers: Rolling SVM and Rolling NSC. Thus, such an ambiguous outcome can be interpreted as valid results, or as a palimpsest, through which the nature of a collaboration can be seen. One author (Ford) seems to have been responsible for setting the plot and drafting the chapters—which is visible in long vectors of mostly content words examined—while the other author (Conrad) seems to have done the actual word weaving (or heavy editing), the trace of which has been left in the function words' usage. In any case, further benchmarks are needed to corroborate the palimpsestic hypothesis.

4 Conclusions

The sequential method as introduced above seems to be an attractive addition to the existing stylometric toolbox. Unlike standard procedures, aimed at assessing style differentiation between discrete text samples, it tries to look inside a text represented as a set of linearly sliced chunks, in order to test their stylistic consistency. Supported by compact visualization, and available, so far, in three variants—Rolling SVM, Rolling NSC, and Rolling Delta—the technique is designed to assess mixed authorship. However, it can be also used as a magnifying glass to inspect works that behave strangely when analyzed using classical stylometric methods: if, say, *Night and Day* by Virginia Woolf does not cluster together with other works by the same author, it might be interesting to see which parts of the novel are more, and which less similar to Woolf's authorial voice.

However, no matter how promising the rolling method is, it has also some drawbacks, at least at the current state of development. The most dangerous one is the risk of accidental false positives, as could be observed in the case of *Roman the la Rose* and *Queen Sophia's Bible*. Arguably, this is due to an inherent characteristic of any written text rather than a pitfall of the method itself: even if sliced into considerably long samples, some local authorial idiosyncrasies cannot be avoided. In this regard, the modernists, and generally the authors experimenting

with style, will probably be more difficult to pinpoint than stylistically consistent realistic literature.

In standard supervised machine learning, the stage of attributing usually is, or at least should be, accompanied by a cross-validation procedure. This issue, as an important addition to the rolling classification techniques, will be discussed in a sequel of this article.

Appendix

The rolling stylometry technique as introduced above is supported by the R package 'stylo' (ver. 0.5.8), through the function *rolling.classify()*. However, since the function is not supplemented by GUI yet, it might look nonintuitive. This Appendix provides a concise step-by-step how-to explaining its usage. To reproduce the experiments discussed in this document, one should use the following procedure:

- (1) Install the package stylo in the version 0.5.8 or later. In R shell, type:

```
install.packages("stylo")
```

- (2) Create a new folder: it will serve as a working space for your experiment. Create two sub-folders named 'test_set' and 'reference_set' (all file names are case sensitive!). Put your disputed text into the 'test_set', put the remaining texts into the 'reference_set'. The setup for the Conrad/Ford case was as follows:

reference_set

```
Conrad_Chance_1913.txt
Conrad_Heart_1899.txt
Conrad_Lord_1900.txt
Conrad_Nigger_1897.txt
Conrad_Victory_1915.txt
Conrad_Western_1911.txt
Ford_Apollo_1908.txt
Ford_Benefactor_1905.txt
Ford_Girl_1907.txt
Ford_Nancy_1911.txt
Ford_Seal_1907.txt
Ford_Soldier_1915.txt
```

test_set

```
Conford_Inheritors_1901.txt
```

The complete setup for *Roman de la Rose* case can be downloaded from the Computational Stylistic Group website <<https://sites.google.com/site/computationalstylistics/>>.

- (3) Load the library ‘stylo’, set your working directory:

```
library(stylo)
setwd("path/to/the/folder/containing/
two/subcorpora")
```

- (4) Optional but important: read the help page for the function *rolling.classify()*:

```
help(rolling.classify)
```

- (5) Run the function *rolling.classify()*, use as many arguments as needed:

```
# Fig. 5
rolling.classify(write.png.file=
TRUE, classification.method="svm",
mfw=100, training.set.sampling=
"normal.sampling", slice.size=5000,
slice.overlap=4500)
# Fig. 6
rolling.classify(write.png.file=
TRUE, classification.method="svm",
mfw=500, training.set.sampling=
"normal.sampling", slice.size=10000,
slice.overlap=9000)
```

The vertical dashed line that divides the part by Guillaume de Lorris and Jean de Meun is produced by adding the word ‘xmilestone’ into the input text, after the line 4,058 (i.e. after *ca.* 50,000 words). One can add as many milestones as needed; they will be reproduced in the final plot and labeled automatically using lowercase roman letters.

```
# Fig. 7
rolling.classify(write.png.file=
TRUE, classification.method="nsc",
mfw=100, training.set.sampling=
"normal.sampling", slice.size=5000,
slice.overlap=4500)
# Fig. 8
rolling.classify(write.png.file=
TRUE, classification.method="nsc",
```

```
mfw=500, training.set.sampling=
"normal.sampling", slice.size=5000,
slice.overlap=4500)
# Fig. 2
rolling.classify(write.png.file=
TRUE, classification.method="delta",
mfw=1000)
# Fig. 3
rolling.classify(write.png.file=
TRUE, classification.method="delta",
mfw=500)
# Fig. 4
rolling.classify(write.png.file=
TRUE, classification.method="delta",
mfw=100)
```

Black-and-white variants of the above plots can be produced using an additional parameter:

```
# Fig. 4a, black-and-white version
rolling.classify(write.png.file=
TRUE, classification.method="delta",
mfw=100, colors.on.graphs=
"greyscale")
```

The source code of the package ‘stylo’, including the newly added function *rolling.classify()*, can be downloaded from the GitHub repository: <<https://github.com/computationalstylistics/stylo>>.

References

- Burrows, J. (2010). Never say always again: reflections on the numbers game. In McCarty, W. (ed.), *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviors, Products and Institutions*. Cambridge: Open Book Publishers, pp. 13–36.
- Burrows, J. F. (2002a). “Delta”: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267–87.
- Burrows, J. F. (2002b). The Englishing of Juvenal: computational stylistics and transtated texts. *Style*, 36: 677–99.
- Deptuchowa, E. (2008). *Odpowiedniki czeskiego aorystu w Biblii królowej Zofii*. Kraków: Lexis.
- Eder, M. (2011). Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 6: 99–114. <http://www.wuj.pl/page,art,artid,1923.html>.

- Eder, M.** (2015). Does size matter? Authorship attribution, short samples, big problem. *Digital Scholarship in the Humanities* **30**. Advance Access published 14.11.2013; DOI:10.1093/llc/fqt066.
- Eder, M., Kestemont, M., and Rybicki, J.** (2013). Stylometry with R: a suite of tools. In *Digital Humanities 2013: Conference Abstracts*. Lincoln, NE: University of Nebraska-Lincoln, pp. 487–9.
- Herdan, G.** (1966). *The Advanced Theory of Language as Choice and Chance*. Berlin: Springer.
- Hirst, G. and Feiguina, O.** (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, **22**(4): 405–17.
- Hoover, D. L.** (2004). Testing Burrows's delta. *Literary and Linguistic Computing*, **19**(4): 453–75.
- Hoover, D. L.** (2011). "The Tutor's Story": a case study of mixed authorship. In *Digital Humanities: Conference Abstracts*. Stanford, CA: Stanford University, pp. 149–51.
- Jockers, M. L. and Witten, D. M.** (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, **25**(2): 215–23.
- Kestemont, M.** (2010). Velthem et al. A stylometric analysis of the rhyme words in the account of the Battle of the Golden Spurs in the fifth part of the Spiegel histor-iael. *Queeste*, **17**: 1–34.
- Kestemont, M. and van Dalen-Oskam, K.** (2009). Predicting the past: memory-based copyist and author discrimination in medieval epics. In *Proceedings of the 21st Benelux Conference of Artificial Intelligence (BNAIC) 2009*. Eindhoven, pp. 121–8.
- Khmelev, D. V. and Teedie, F.** (2001). Using Markov chains for identification of writers. *Literary and Linguistic Computing*, **16**(3): 299–307.
- Koppel, M., Schler, J., and Argamon, S.** (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60**(1): 9–26.
- Kyas, V.** (1997). *Česká bible v dějinách národního písemnictví*. Vyšehrad: Nakladatelství Vyšehrad.
- Markov, A. A.** (2006 [1913]). An example of statistical investigation of the text "Eugene Onegin" concerning the connection of samples in chains, trans. into English by G. Custance and D. Link. *Science in Context*, **19**(4): 591–600.
- Marteau, P. (ed.)** (1878). *Le Roman de la rose, par Guillaume de Lorris et Jean de Meung*. Éd. accompagnée d'une traduction en vers, précédée d'une introduction, notices historiques et critiques; suivie de notes et d'un glossaire par Pierre Marteau. Vol. 1–4. Orléans: H. Herluison.
- Pawłowski, A.** (1999). Language in the line vs. language in the mass: on the efficiency of sequential modelling in the analysis of rhythm. *Journal of Quantitative Linguistics*, **6**(1): 70–7.
- Pawłowski, A. and Eder, M.** (2001). Quantity or stress? Sequential analysis of Latin prosody. *Journal of Quantitative Linguistics*, **8**(1): 81–97.
- Petruszewycz, M.** (1981). *Les Chaînes de Markov Dans le Domaine Linguistique*. Genève: Slatkine.
- Rybicki, J.** (2012). The great mystery of the (almost) invisible translator. In Oakes, M. P. and Ji, M. (eds), *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins, pp. 231–48.
- Rybicki, J. and Heydel, M.** (2013). The stylistics and stylometry of collaborative translation: Woolf's "Night and Day" in Polish. *Literary and Linguistic Computing*, **28**(4): 708–17.
- Rybicki, J., Kestemont, M., and Hoover, D.** (2014). Collaborative authorship: Conrad, Ford and rolling delta. *Literary and Linguistic Computing*, **29**(3): 422–31.
- Stamatatos, E.** (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3): 538–56.
- Tabata, T.** (2014). Stylometry of collaborations: Dickens, Collins and their collaborative writings. In *Digital Humanities 2014: Conference Abstracts*. Lausanne: EPFL-UNIL, pp. 378–80.
- Thaisen, J.** (2012). A probabilistic analysis of a Middle English text. In Nelson, B. and Terras, M. (eds), *Digitizing Medieval and Early Modern Material Culture*. Tempe, Arizona: Center for Medieval and Renaissance Studies, pp. 171–200.
- Twardzik, W. (ed.)** (2006). *Biblioteka zabytków polskiego piśmiennictwa średniowiecznego*. Kraków: IJP PAN. [a corpus of Old-Polish on a DVD].
- Urbańczyk, S.** (1933). Jeden czy kilku tłumaczy Biblii szarospatackiej. *Slavia Occidentalis*, **13**: 25–8.
- van Dalen-Oskam, K. and van Zundert, J.** (2007). Delta for Middle Dutch: author and copyist distinction in Walewein. *Literary and Linguistic Computing*, **22**: 345–62.
- Wanicowa, Z.** (2009). *Ignota, dubia, reperta. Czytać i rozumieć staropolszczyznę*. Kraków: Lexis.