DHQ: Digital Humanities Quarterly
Preview
2023
Volume 17 Number 4

Problems of Authorship Classification: Recognising the Author Style or a Book?

František Válek <frantisek_dot_valek_at_upce_dot_cz>, National Library of the Czech Republic; Department of Philosophy and Religious Studies, University of Pardubice; Institute of Ancient Near Eastern Studies, Charles University
https://orcid.org/0000-0003-1449-004X
Jan Hajič, Jr. <hajicj_at_ufal_dot_mff_dot_cuni_dot_cz>, National Library of the Czech Republic; Masaryk Institute and Archive, Czech Academy of Sciences; Institute of Formal and Applied Linguistics, Charles University
https://orcid.org/0000-0002-9207-567X

Abstract

The presented article proposes that one of the problems regarding authorship attribution tasks is the attribution of a specific book rather than the author. This often leads to overestimated reported performance. This problem is in general connected to the dataset construction and more specifically to the train-test data split. Using a heavily delexicalized and diverse dataset of Czech authors and basic LinearSVC classifiers, we designed a three-step experiment setting to explore book versus author attribution effects. First, the authorship attribution task is performed on a dataset split to train and test data segments across books. Second, the same task is performed on a dataset where individual books are used wholly either for training or testing. Expectedly, this leads to poorer results. In the third step, we do not attribute book segments to authors but to books themselves. This step reveals that there is a general tendency towards attributing to a specific book rather than to different books of the same author. The results indicate that authors who show a higher inner confusion among their works (i.e., the model attributes their works to other works of theirs) tend to perform better in the task of attribution of an unseen book.

Introduction

Authorship attribution using machine learning is a fertile area of digital literary scholarship[1] to such an extent that it has its own software package within the R programming language ecosystem.[2] Feature design is the most interesting sub-problem in terms of classification performance; very recently, Robert Gorman has achieved impressive results with morphosyntactic instead of lexical features, also for very short segments [Gorman 2022]. However, we explore a different aspect of the problem: that of experiment design and, by implication, dataset design. Issues of dataset imbalance, genre consistency and dataset sizes (both in terms of the total number of tokens and texts, as well as the number of authors) have been discussed, for instance, in the works of Efstathios Stamatatos and Kim Luyckx, but what has received comparatively little attention is stylistic variation among the works of a single author [Stamatatos 2009] [Luyckx 2011]. 1

In this article, we show that in the domain of long-form literary works, stylistic variation between individual works of the same author is a significant factor that should be reflected in the dataset and experiment design. Assuming that we are interested in capturing features of authorial styles that transcend the boundaries of their known works (especially to attribute texts with unclear authorship, such as in the seminal study of *The Federalist Papers* by Frederick Mosteller and David L. Wallace), a test set that includes different segments of works that have been previously used during training will significantly overestimate the system's accuracy for unseen texts and therefore overestimate the system's ability to characterise authorial style, as opposed to the styles of individual works [Mosteller and Wallace 1964]. The extent of this 2 overestimation differs significantly among individual authors; some have a more consistent style than others.

Our findings are applicable when we are interested in attributing texts to authors *despite* their stylistic inconsistencies. This may not always be the case – an authorship attribution system might be used for other purposes, such as to find which observable linguistic features are responsible for stylistic distinctions, where the classification task is merely a proxy for the true goal (which would then be achieved through feature selection). Note also that our findings only apply to classifying texts that are long enough to be processed by segments. This is often the case with applications in the study of literature, less so in attributing authorship of short texts such as emails or tweets. 3

The contribution of our article can be summarised thus: the stylistic variation between individual books of an author is significant enough to affect state-of-the-art authorship classification system performance. Thus, to credibly claim a certain level of ability to classify the style of an *author* as opposed to the style of individual *books*, the evaluation set should contain whole books not seen during training. We believe these findings are useful, first of all, to authorship attribution system designers, as we quantify the extent to which stylistic variation among books matters to the classifier. Thus, we provide a guideline for evaluation design so that system performance is not overestimated. Second, we believe our findings are useful to the literary scholar selecting a classification system to inform judgments about the authorship of unattributed texts, as our findings show that systems that do not test on unseen books cannot be trusted to perform as well as their evaluation results indicate. 4

It should be noted that we did *not* aim to maximise the classification accuracy beyond a reasonable fraction of the state-of-the-art [Tyo, Dhingra, and Lipton 2022]. We used the standard, state-of-the-art Support Vector Machine (SVM) classifier and performed a brief hyperparameter search over possible settings. 5

Crucially, as in [Gorman 2022], we perform *delexicalisation*. This is a step that replaces words (primarily autosemantic words, such as nouns, adjectives, verbs, and adverbs) with just their morphosyntactic properties so that the topics and contents of individual books do not artificially inflate the stylistic differences between individual books. Imagine that the same author wrote one novel from a 19th-century farm environment and one from a Great War factory. While the style in terms of linguistic choices in both books may be very similar, the vastly different content and, therefore, vocabulary would make it difficult to identify the factory book as being written by the same author as the farm book. The process of delexicalisation filters out such content-related confounding factors when focusing on author style detection: names of characters and places, characteristic objects (such as the presence of automobiles or wireless communication), genres (such as realist or anarchist perspectives on social conditions vs. detective stories or gothic fiction), and environments (urban vs. rural, wartime vs. peacetime conditions) and helps to avoid confusion of authors dealing with the same topics or writing about the same geographical areas. On the other hand, one must be aware that delexicalisation implicitly restricts the definition of “authorial style” by excluding vocabulary choices (of autosemantic words) and some elements of register (such as informality expressed in Czech by orthography of word endings). Some aspects of the author style are therefore lost in delexicalisation. However, we consider it more important to remove confounding factors that can identify specific books (and therefore authors), which can hardly be considered elements of author style. Given that we are attempting to explore the extent to which author style varies between books, we want to avoid leveraging trivial sources of this variation. 6

It should be noted that the use of non-lexical features is by no means rare in computational stylometry [Swain, Mishra, and Sindhu 2017]. We use the UDPipe morphosyntactic feature extractor to extract morphological features [Nivre 2015]. (Notably, while [Gorman 2022] uses UDpipe features as well, they combine them in a more sophisticated manner and achieve better absolute accuracies, especially for shorter segments.) 7

In the rest of the article, we first introduce our dataset and specify pre-processing and hyperparameter search procedures and results. Then, we demonstrate our findings in three experimental steps. First, we establish the baseline accuracies for a system that does not distinguish between training and test books. Next, we show how results change once specific books are set aside for testing. Finally, we show how the tension between author and book style is distributed across the dataset. 8

The following visualisation summarises our findings. The columns show results for different segmentation lengths (in 9 tokens, see below). The first part shows the results of our classifiers when each of the books is split into train and test segments. Here, a different selection of test segments does not significantly influence the performance. The following rows show the results when the train-test split is not done across all books, but one book for each author is left out of training and used for testing. With a random selection, five runs were performed. The performance correlates significantly with segment lengths and is also greatly influenced by the test-book selection.

Train and Test Across All Books of the Dataset **(Experiments: Step 1)**					
Segment Length	**s-1000**	**s-500**	**s-200**	**s-100**	**s-50**
Full Dataset	0.96	0.90	0.73	0.58	0.42
Validation	0.96	0.91	0.74	0.58	0.42
Train Books vs. Test Books **(Experiments: Step 2)**					
Set 1	0.86	0.80	0.62	0.44	0.29
Set 2	0.86	0.78	0.58	0.38	0.23
Set 3	0.90	0.82	0.63	0.42	0.26
Set 4	0.77	0.69	0.51	0.35	0.23
Set 5	0.92	0.85	0.66	0.47	0.33

Table 1. Summary table of results.

Dataset

10 Our dataset consists of 210 books (written in Czech) by 23 authors (for a full overview, see the table in the appendix). The authors were chosen from the late 19th and early 20th centuries to avoid differences in the written form of the Czech language due to chronological development in standardisation. This limited timeframe reduces differences in style stemming from the varied periods of origin. The dataset is far from being balanced: for each author, we have chosen a different number of books (ranging from 4 books by Č. Slepánek to 18 books by K. Čapek) of varying lengths (the shortest book consists of only 6,004 tokens while the longest contains 300,021 tokens). In addition, even though novels dominantly prevail, the genres vary across the dataset. See the appendix for a detailed overview of the dataset.

11 Such a diverse and unbalanced nature dataset may not be ideal for machine learning (ML) experiments, but it reflects the reality of library collections and the issues with authorship attribution. There are several features of our dataset that can be contrasted with the dataset of [Gorman 2022] and show that what Gorman presents as a difficult problem must be problematised even further.[3]

12 First, we have included several books by each of the authors. Therefore, our dataset has the potential to demonstrate whether different authors change their style across their works. As is discussed below, our experiments have shown that some authors are more consistent across their work, allowing us to accurately attribute to them a book which has never been seen within the training process, while other authors vary their style to such an extent that attributing an unseen book to them is almost impossible. In these cases, when trained and tested across the dataset, we are actually attributing the style of texts to individual books rather than the authors.

13 Second, the books we have chosen vary greatly in length. [Gorman 2022] has chosen works that include at least 20,000 tokens. In our dataset, we have 17 books that do not reach this limit, but we compensate for this by including more books for each of the authors, so there are significantly more than 20,000 tokens for each author, ranging from 174,115 tokens for Č. Slepánek) to 1,512,167 tokens for A. Jirásek.

14 Finally, our dataset includes mainly prose (mostly literary, but also journalistic and scholarly), a few works of drama, and one item of poetry. This further complicates the problem mentioned in the previous paragraph. While [Gorman 2022] is right that varying genres may lead to the confounding of genres with author styles, we believe that we can learn something interesting from including such data. In the end, our experiments have shown that author style remains

partially preserved across genres. Expanding the dataset with works of drama and poetry may be fruitful in the future, but this must be done hand-in-hand with an expansion of author selection (as the selected authors are predominantly novelists).

Data Preparation

Data Cleaning

The raw data we have at our disposal are scanned books that have been processed with optical character recognition (OCR).[4]Therefore, some cleaning was necessary. Basic automatised data cleaning was performed[5], followed by a manual clearing of junk data such as imprints, tables of contents, forewords, afterwords, and endnotes. Finally, hyphenated words were restored across line boundaries and page boundaries. For the sake of keeping the pipeline simple, we did not fix OCR errors; they could, however, be mitigated using, for example, the LINDAT Korektor service.[6] 15

Segments and Train-Test Split

For training and testing authorship detection, we must split the books into shorter segments. For testing, we need a sufficient number of test samples to provide meaningful accuracy estimates. For training, this is necessary to provide a sufficient number of data points while keeping the segments long enough to provide meaningful estimates of the relationship between feature distributions and segment authors. 16

We split the dataset into segments of 1000, 500, 200, 100, and 50 tokens as data points for classification experiments, denoted s-1000, s-500, and so on. Because we want the dataset to allow us to investigate how authorial style is expressed through other than lexical choices, including potentially syntactic features (although we do not use those in this work), we decided only to draw segment boundaries at the sentence level. Thus, these segment lengths represent the *average* segment lengths because sentences occur in lengths that do not sum exactly to the desired multiple of 50. We discarded end-of-book segments if they were shorter than half the target segment length. To maintain a consistent training and test set so that results are directly comparable between segment lengths, we first built the s-1000 segmentation, assigned these segments to training and test sets, and then obtained the shorter segmentations by splitting the s-1000 segments, rather than re-segmenting the entire books. This ensures that each test segment in the shorter segmentations is a subset of a test segment in s-1000, and each training segment in shorter segmentations is a subset of a training segment in s-1000, maintaining the same content of the test and training sets across different segment lengths. (Note that this is a dataset design choice, not an experiment design, with the primary aim of enabling a direct comparison to the results presented in Experiment and Results, Step 1.) 17

Specifically, we have pre-split the data into “train” (60%), “development” (20%), and “test” (20%) segments in order to make future direct comparisons to our results with this dataset straightforward.[7] However, as we have not made any attempts at optimising the classifiers and instead used their default settings (see below), unless stated otherwise, the “training” set for all our experiments consists of both the “train” and “development” subsets of the dataset. 18

Delexicalisation

As stated above, we have delexicalised the dataset using the publicly available Application Programming Interface (API) of UDPipe at LINDAT/CLARIAH.[8] The UDPipe service performs canonical tokenisation and outputs a set of extracted features for each token. For the purpose of this article, we have applied delexicalisation that replaces all of the autosemantic words[9] by their part-of-speech tag[10] and all other words by their lemmas. 19

In contrast, [Gorman 2022] has provided a more nuanced and sophisticated approach to delexicalisation that indeed seems much more fruitful. In the future, combining the variety of experiments presented in this paper and enhanced classifiers using more sophisticated forms of delexicalisation may yield more significant results. Nonetheless, while not being the state-of-the-art approach, using POS and lemmatisation is well established in the authorship attribution field [Swain, Mishra, and Sindhu 2017]. 20

In addition to using the above mentioned form of delexicalisation, we have performed a variety of delexicalisations for a smaller subset of 6 authors (30 books) to explore the effects of different levels of delexicalisation on the performance of various classifiers (see below). Still, these forms of delexicalisation do not reach the complexity of the approach utilized by [Gorman 2022]. 21

Hyperparameter Search: Authorship Classification at Varying Levels of Delexicalisation

To set reasonable parameters for the pipeline, we conducted a series of experiments, working as a kind of grid search, to explore the results of different classifiers in relation to different levels of delexicalisation. This is a "lightweight" hyperparameter search that helps us find a model and pre-processing settings such that we do not work with an unnecessarily underperforming setup, rather than finding an optimal setup for the dataset. 22

Because these experiments are essentially a grid search, we have selected only a subset of our data, consisting of six authors (5 books per each, 30 in total): A. Stašek, J. Neruda, J. Arbes, K. Klostermann, F. X. Šalda, and T. G. Masaryk. [11] 23

For these experiments, we chose one of 10 different levels of delexicalisation. All of these pre-processings have been segmented in the same way as the larger dataset used for the rest of the experiments (1000, 500, 200, 100, and 50 tokens). 24

The different modes of delexicalisation were abbreviated as "r-codes", from r-04 to r-13.[12] The baseline where no delexicalisation was applied is r-04. Delexicalisations based on UDPipe are applied in r-05 through r-09.[13] We also applied NameTag 2 [Straková, Straka, and Hajič 2019] to replace named entities with tags specifying only the type of the named entity in r-10 through r-13.[14] The full list of delexicalisation settings we explored is as follows: 25

- r-04: No delexicalisation (baseline) — original word forms are used
- r-05: Lemmatisation — lemmas used instead of word forms
- r-06: Part-of-speech tags for all words
- r-07: Morphological tags for all words
- r-08: Part-of-speech tags for autosemantic words, others lemmatised
- r-09: Morphological tags for autosemantic words, others lemmatised
- r-10: NameTag tags for recognised named entities, others with original word forms
- r-11: NameTag tags for recognised named entities, others lemmatised
- r-12: NameTag tags for recognised named entities, part-of-speech tags for autosemantic words, others lemmatised
- r-13 NameTag tags for recognised named entities, morphological tags for autosemantic words, others lemmatised

We then conducted a series of experiments across all of these levels of delexicalisation as well as across different segmentations. We have trained the following standard classifiers used in authorship classification [Savoy 2020, chap. 6], using the default implementations and hyperparameter settings in the scikit-learn library (https://scikit-learn.org/) [Pedregosa et al. 2011]: 26

- Naive Bayes (sklearn.naive_bayes.MultinomialNB)
- C-Support Vector Classification (sklearn.svm.SVC)
- Linear Support Vector Classification (sklearn.svm.LinearSVC)
- K-Nearest Neighbours (sklearn.neighbors.KNeighborsClassifier)
- Stochastic Gradient Descent (sklearn.linear_model.SGDClassifier)
- Decision Tree (sklearn.tree.DecisionTreeClassifier)

We used the same feature extraction settings for each (sklearn.feature_extraction.text.CountVectorizer). The only 27 adjusted setting was word n-gram size, set to unigrams, bigrams, and trigrams (vectorizer = CountVectorizer(ngram_range=(n_min, n_max))). The training was performed using only the “train” passages, and the evaluation using only the “test” passages. Because we used the default settings, the “devel” passages were unnecessarily ignored during this phase. We performed multiple runs of trainings and evaluations across the classifiers and pre-processing.

28 As an example, we provide here a table of results representing the accuracy scores of the LinearSVC classifier run across all books with different pre-processing.

Segment Length	**s-1000**	**s-500**	**s-200**	**s-100**	**s-50**
r-04	0.99	0.99	0.97	0.94	0.87
r-05	1.00	0.99	0.97	0.94	0.87
r-06	0.95	0.91	0.79	0.69	0.60
r-07	0.97	0.96	0.90	0.83	0.71
r-08	0.96	0.95	0.86	0.77	0.62
r-09	0.97	0.96	0.89	0.82	0.70
r-10	0.99	0.98	0.96	0.93	0.86
r-11	1.00	0.98	0.97	0.93	0.86
r-12	0.97	0.95	0.88	0.78	0.64
r-13	0.98	0.95	0.89	0.83	0.71

Table 2. Table of results using varying levels of delexicalisation. The higher the number, the better the classification performance.

29 These initial experiments have shown a performance dependency on different levels of delexicalisation. In addition, it has became clear that different classifiers react differently across varying levels of delexicalisation. This also shows us that we are not recognising the author style *per se*, but rather that we are constantly flattening the problem to how the author style is reflected in specific conditions using specific features. Quite unsurprisingly, the performance is highly dependent on segment lengths.

30 Of all the classifiers we experimented with, support vector machines worked best. Therefore, we have decided to use LinearSVC for the rest of the experiments, using the pre-processing r-08. Even though the pre-processing r-08 did not lead to the best performance, it represents a simple and straightforward yet very strong level of delexicalisation that significantly reduces the number of features and conceals content.

Experiments and Results

31 Using the pre-processing/classification pipeline described above, we performed three experimental steps to illustrate the relationship between author and book style in authorship classification:

1. The full dataset (see above; 23 authors, 210 books) was used for a task of authorship attribution with *each book* divided into “train” (80%) and “test” (20%) passages. We reported performance across different lengths of passages. This is an “easy” setting for the classifier.
2. Next, we performed the experiment with the same settings, but this time we built the “test” set by choosing *one book from each author* and adding *all its segments* into the test set. All other books of each author were used in their entirety for training. In this “harder” setting, performance dropped significantly, which is the main point of this paper. Furthermore, classification performance was influenced by the selection of the testing books.
3. Finally, we performed the same experiment as outlined in #1, but instead of classifying by author, we classified segments into individual books. With this experiment, it is possible to discuss further why the test-book selection in Experiment #2 is so influential, as well as to show that some authors are more consistent

in their style (as expressed by the selected features) than others.

We again emphasise that our purpose here is not to reach the best possible classification accuracy but rather to explore the influence of authorial style variation between individual works on the classification accuracy of a “decent” pipeline. Our pre-processing steps, classification models, and final scores are not the focus of our findings. Rather, we are interested in exploring how classification metrics change across different experiments. 32

Step 1. Train and Test Across All Books of the dataset

Having selected the pre-processing and classification pipeline (delexicalisation r-08, using part-of-speech tags instead of autosemantic words and lemmas for functional words, and the LinearSVC classifier), we measured the baseline results when “train” and “test” sets were drawn randomly from all books of each author. 33

In addition to measuring performance on the “train+devel versus test” split, we also tried an alternative “train+test vs. devel” split (which, because we never used the development set for model selection, is essentially just a different partition for cross-validation). The results were almost identical, so we did not consider it necessary to carry out full cross-validation. 34

The following table shows the accuracies of the two experiment runs in comparison with the same setting on a smaller dataset. 35

Across Books	**s-1000**	**s-500**	**s-200**	**s-100**	**s-50**
Full Dataset	0.96	0.91	0.74	0.58	0.42
Full Dataset, Validation	0.96	0.9	0.73	0.58	0.42
Small Dataset (6 Authors)	0.96	0.95	0.86	0.77	0.62

Table 3. Step 1 results table.

The selection of individual segments as testing seems to have only limited influence on the results. The best results were achieved, predictably, using the longest segments of 1000 tokens (96.1/96.4%). Shorter segments significantly lowered the performance.[15] 36

Interestingly, when using the segments of 1000 tokens, the accuracy of the classifier was consistently around the same 96% for both the small, six author dataset used for hyperparameter selection and for the full dataset of 23 authors. However, when shortening the segments, the performance on the larger dataset radically dropped.[16] 37

The results of this first experiment serve as a baseline for the second, where we show how setting aside specific books changes the results. 38

Step 2. “Train” Books Versus “Test” Books

We believe that the above-mentioned experiments are relatively simple ML-based author attribution tasks. In our opinion, the experimental settings that use the same books for training and testing, such as [Gorman 2022] or [Benotto 2021], are biased in their reported performance. After all, the capability to recognise the author of an unseen and unattributed text is one of the main research objectives within the field of authorship attribution (while certainly not being the only goal [Swain, Mishra, and Sindhu 2017]). 39

Therefore, we have further expanded the experimental scenario to address real-life problem: recognising a book (or rather its parts) that has never been seen in the system building process (see the appendix for detailed information). The experiments discussed below reveal that selecting a test book from the available corpus heavily influences the reported performance of the classifier. 40

We randomly selected five sets of testing books such that each set contained one book from each author and no book 41 was in two testing sets, except for *Svědomí Lidových novin, čili, Jak bylo po léta v českém tisku štváno lživě proti mně* (a-08.b-03) in sets 1 and 5, because the dataset contains only four books by Č. Slepánek. We ran the same classification experiment and reported results across segment sizes. The results are reported in the following table.

Book-Based	**s-1000**	**s-500**	**s-200**	**s-100**	**s-50**
Set 1	0.86	0.80	0.62	0.44	0.29
Set 2	0.86	0.78	0.58	0.38	0.22
Set 3	0.90	0.82	0.63	0.42	0.26
Set 4	0.77	0.69	0.51	0.35	0.23
Set 5	0.92	0.85	0.66	0.47	0.22
Average	0.86	0.79	0.60	0.41	0.24
Cf. Step 1 (Across Books)	0.96	0.91	0.74	0.58	0.42

Table 4. Step 2 results table.

The results show a drop of 0.04 to 0.23, with 0.10 being the average deterioration of classification accuracy. In the case of the easiest s-1000 and s-500 settings, this means more than a three-fold increase in error. Furthermore, Set 4 shows that a random selection of testing books can make this difference much larger. 42

We note that performing five-fold, cross-validation with 23-book test sets rather than 210-fold, leave-one-out, cross-validation on individual books had little bearing on these results while being significantly more expedient despite the classifier performance on each testing book being further influenced by the choice of the other 22 testing books in each fold. We chose the five worst performing outliers and five high-performing books and performed leave-one-out experiments with these. We found that the leave-one-out results were, in fact, worse by 0.5% on average (when disregarding 3 books that had their authors classified perfectly in the 5-fold and leave-one-out settings both), with the leave-one-out accuracy ranging from 7.6% higher (a-03 test book from Set 4) to 7% lower (a-15 test book from Set 1). 43

Compared to the leave-one-out setting, the effect of removing books potentially helpful for identifying an author from the training set was roughly cancelled out by the effect of introducing potentially confounding books to the training set. As a result, while the estimates for individual books did likely have a somewhat higher variance, our main finding that accuracy dropped significantly overall in this setting was not affected. Furthermore, the accuracies of items of the highest significance for further analysis — outliers in both directions — seem to have been affected by less than 10%, which does not materially affect the selection of books that are significantly harder or easier to classify by author than the average. Thus, our analytical attention is directed to the same items that a leave-one-out experiment design would point towards. [17] 44

These experiments show that there is little to be gained by performing the remaining 199 leave-one-out experiments over the 5-fold scheme. We attribute this consistency between the lower-variance, leave-one-out setting and the 5-fold setting to the size of the dataset: at these scales, leave-one-out cross-validation schemes no longer provide a less biased estimate of aggregate statistics, and the effect of inclusion of individual items into the training set is not as pronounced. Note also that although in our 5-fold cross-validation the folds differed by 46/210 books, the results for each book within a fold were computed on a perfectly identical training set and thus are perfectly comparable, while in the leave-one-out setting, no two training sets are the same. 45

A possible systematic confounding factor for the drop in average performance could be the irregularity of training set sizes introduced by setting aside random entire books for testing, as book lengths vary greatly. As opposed to the “Across Books” setting from Step 1, here we do not have the same “train”:“test” token ratio. The obvious question then arises: is the model performance dependent on the “train”:“test” ratio, “test” or “train” token absolute count, both, or neither? The following table shows data from experiments (for s-1000 segments). The indicated “test ratio” is the ratio of “test” tokens to all tokens. Asterisks indicate the experiments where drama or poetry were used as the test book (* = drama; ** = poetry). 46

	All Tokens	Set 1 Rest Ratio Accuracy	Set 2	Set 3	Set 4	Set 5
a-01	1,044,186	7.76% 0.97	13.50% 0.91	6.38% 1.00	8.81% 0.97	8.79% 0.82
a-02	364,582	4.05% 1.00	17.00% 0.53	21.00% 0.94	24.43% 0.88	33.51% 0.83
a-03	1,197,470	6.10% 1.00	3.34% 1.00	7.43% 0.93	13.62% 0.20	10.44% 0.99
a-04	371,909	25.74% 0.83	27.00% 0.90	23.46% 0.91	4.94% 0.67	18.86% 0.97
a-05	285,386	22.12% 0.95	39.89% 0.81	28.83% 0.85	3.55% 0.90	5.90% 0.94
a-06	299,530	63.53% 0.64	8.56% 0.96	3.89% 0.83	2.93% 0.78	21.09% 0.83
a-07	1,512,167	14.22% 0.98	4.96% 0.91	1.79% 0.70	11.90% 0.99	9.66% 0.99
a-08	174,115	3.45% 0.00	63.19% 0.13	10.37% 0.67	22.99% 0.10	3.45% 0.00
a-09	374,104	10.69% 0.45	*3.21% *0.42	12.57% 0.91	19.79% 0.24	24.06% 0.80
a-10	514,131	19.26% 0.97	14.20% 0.79	13.42% 1.00	3.70% 0.53	13.81% 0.93
a-11	715,093	11.33% 0.86	5.73% 0.78	6.43% 1.00	16.36% 0.74	9.93% 0.93
a-12	241,111	14.52% 0.66	*8.71% *0.86	*6.23% *0.53	*14.53% *0.89	10.39% 0.32
a-13	417,080	8.16% 0.62	16.78% 0.89	14.15% 0.90	16.54% 1.00	14.39% 0.92
a-14	731,207	5.88% 1.00	4.10% 1.00	12.45% 1.00	15.73% 0.96	7.80% 1.00
a-15	785,198	5.35% 0.24	*3.06% *0.96	10.57% 0.84	*2.93% *0.83	*3.18% *0.84
a-16	1,099,103	12.46% 0.90	27.30% 0.96	7.01% 0.99	4.00% 0.93	11.46% 0.95
a-17	614,032	4.40% 1.00	23.45% 0.86	37.46% 0.96	*3.26% *0.75	17.43% 1.00
a-18	819,145	10.74% 0.98	5.98% 0.98	15.14% 0.96	7.33% 1.00	15.63% 0.95
a-19	765,197	2.75% 0.86	12.81% 1.00	9.42% 0.99	9.15% 0.89	8.49% 0.97
a-20	1,137,133	3.52% 1.00	6.16% 0.84	4.13% 0.98	2.11% 1.00	15.04% 0.99
a-21	703,121	5.41% 1.00	12.38% 0.93	24.18% 0.55	**2.42% **0.65	8.25% 0.98
a-22	618,089	6.15% 0.79	5.50% 1.00	5.34% 0.94	2.91% 0.56	17.64% 0.88
a-23	683,108	9.66% 1.00	25.92% 0.98	9.37% 1.00	16.84% 1.00	9.08% 0.98
Average (Per Author)	672,443	12.05% 0.81	15.32% 0.84	12.65% 0.89	10.03% 0.76	12.97% 0.86

Full Performance	672,443	0.86	0.86	0.90	0.77	0.92

Table 5. Step 2 results table for individual authors, showing test ratio and accuracy across the five test book sets. * = drama; ** = poetry.

The following table shows the correlation coefficients of the authors' accuracies to the “train”:“test” ratio, full token count, “test” token count, and “train” token count: 47

	Set 1	Set 2	Set 3	Set 4	Set 5
Train Ratio	-0.19474	-0.01766	-0.01495	-0.18115	0.176758
Full Token Count	0.356384	0.390591	0.234502	0.259228	0.502111
Test Token Count	0.091817	0.217205	0.126774	0.099459	0.510302
Train Token Count	0.358298	0.361026	0.217488	0.268124	0.477041

Table 6. Correlation coefficients of the authors' accuracies to the “train”:“test” ration, full token count, “test” token count, and “train” token count.

These data show that while there are some tendencies in the correlations, in general, these correlations are unstable, and the most significant feature that influences classification accuracy is the selection of the test books. At the same time, there does seem to be a minimum number of tokens necessary in order for the model to perform well. This is visible with Č. Slepánek (a-08). He has the lowest number of tokens, as well as books, and usually performs the worst, except for set 3, where two other authors (a-12 and a-21) perform worse. There is likely not enough data to be trained on, and at the same time, there are only a few test passages (only 3 in the case of sets 1 and 5), so the model has few chances to accurately predict his authorship of a segment, further increasing the variance of the result. 48

On the other end of the data size spectrum, K. Sabina (a-07), who has the highest number of tokens, performs very well but not the best, and his worst performing test book is the shortest of the five — the one with the smallest impact on training data size. A high number of training tokens by itself apparently does not ensure stable performance. Another example may be given in Set 4, a-03 (J. Arbes). Even though the token count is very high, the performance is only 0.20. This deviation is, however, easily explained once the test book is consulted and compared to the rest of his works. In this case, a-03.b-04 (*Persekuce lidu českého v letech 1869-1873*) has been used for testing. In contrast to Arbes' more typical short novels, this book is a work of his journalism career. A similar influence may be observed in the case of V. Hálek (a-21) in Set 3, as the work used for testing (*Fejetony*) is also journalistic. 49

This further opens the question of the influence of genres on the performance of the models. In general, there are journalistic works counted among the prose, and we may also point to several cases of drama or poetry. The works of drama were used for testing in the case of four authors on eight instances[18] and a work of poetry in one case[19]. The influence of genre is not that significant for K. Čapek (a-15, sets 2, 4, 5) or J. Vrchlický (a-12, sets 2, 3, 4), likely because, in their cases, there are several books of drama that provide a sufficient base for testing. 50

On the other hand, for V. Hálek (a-21, Set 4) and K. Sabina (a-17, Set 4), the deviance in genre resulted in a significant drop in performance, probably because there is no training data for support. However, even though the performance significantly dropped in these cases, it was still much higher than a random baseline. Furthermore, other authors who write consistently in one genre performed much worse. 51

There are several other cases where the influence of the selected test book can be well explained. For example, V. Hálek (a-21) shows a 1.00 accuracy in Set 1. A simple look at the dataset does not explain such a success. However, the work *Na statku a v chaloupce* (a-21.b-09) is a short story that is also included in *Kresby křídou i tuší* (a-21.b-10), which was used for training. Such overlaps in datasets are easily created when based on real-life library scenarios. 52

We believe that the data and discussion presented here clearly illustrate the problem and influence of the test-book selection. In addition, we can see that reporting the overall statistics of authorship classification performance can cover 53 up significant specific high-variance issues that come up in more detailed analysis.

Step 3. Books as Targets

Our third experiment aimed to discover the structure of stylistic similarity within individual authors. As hinted by the large differences between cross-validation runs, the stylistic differences among individual books of an author vary significantly. We are interested in the characteristics of these dissimilarities. 54

To expose these characteristics, we ran the classification pipeline with the 210 individual books (instead of the 23 authors) as output classes and observed misclassification patterns. If an author's style is highly consistent, we would expect (thanks to delexicalisation) that segments from one of their books would be easily misclassified as segments from their other books, especially in the “easy” s-1000 segmentation setting where the confusion between authors was minimal. More generally, we have three kinds of possible results for the classification of a segment: (a) the correct book, (b) a wrong book by the correct author, or (c) a book by a different author. If we assume that misclassification is a good proxy for similarity (which we examine later), then we can define the following: 55

- The greater the ratio of b / (a + b), the more consistent an author's style is.
- The greater the ratio of a / (b + c), the more inconsistent an author's style.
- The greater the ratio of (a + b) / (a + b + c), the more distinctive an author is.

We have split each book into 80% of training and 20% of test segments (see above). In this step, two sets of experiments were run. In the first, only books with over 20,000 tokens were used to ensure decent model training (see, for example, [Gorman 2022]). In the second, we included all of the books. Here, we report and discuss only the results of s-1000 segments. The following charts show general results for individual authors. The numbers stated are the numbers of segments which have been attributed to the correct book (blue), other books of the same author (yellow), and books of a different author (red). 56

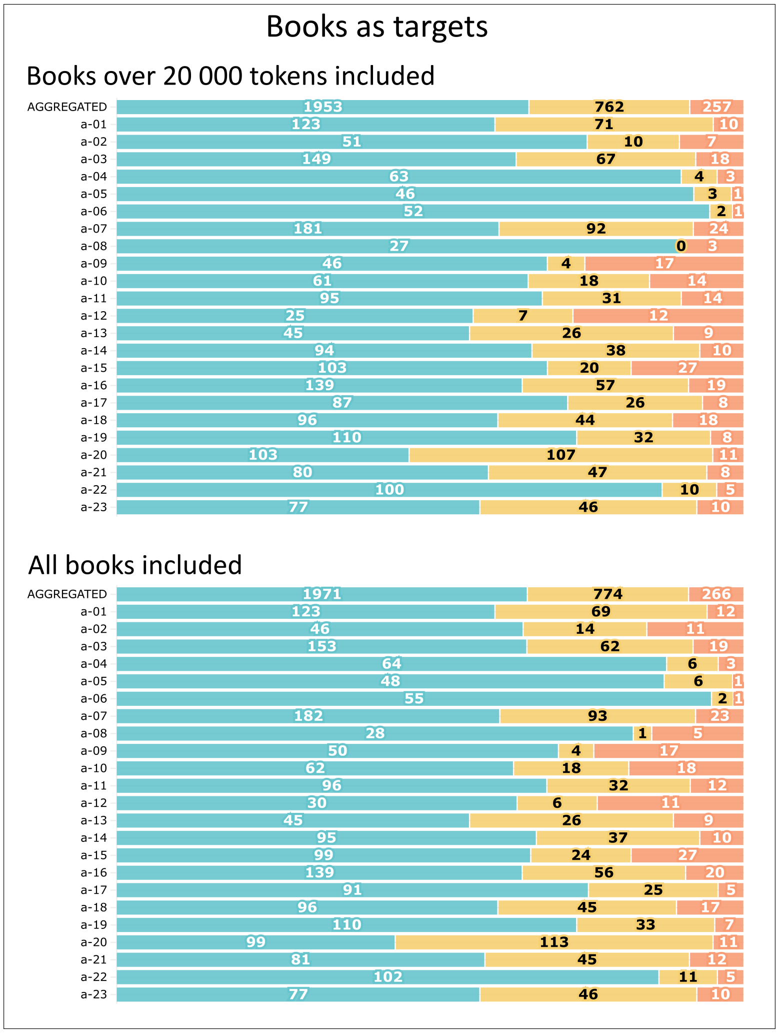

Figure 1. Results of classification of books by individual authors, in numbers of segments attributed to the correct book (blue), other books of the same author (yellow), and books of a different author (red); created using Flourish (https://flourish.studio/, accessed 5 April 2023)

These results may be compared with the experiments performed in Step 2. The author a-20 (S. K. Neumann) can be taken as an author whose style seems consistent across books, whereas author a-06 (T. G. Masaryk) is one whose style seems more book related. In Step 2, a-20 performed better than average, except for Set 2, where the results were slightly below-average (0.84). In that case, the test book was a-20.b-14. This book consists of 70 passages, meaning it has 14 test passages. Ten passages were attributed to the same book, two to other books of the same author, and two to incorrect authors. Author a-06 seems to be more consistent within individual books, but these are only rarely confused with each other. In the experiments of Step 2, a-06 performed best in Set 2 (0.96). The test book selected for this set (a-06.b-05) is the only one that gets confused with the author's other books. Even though there are only 5 test passages in this short work, the results of steps 2 and 3 seem to correlate. The following chart shows the heatmaps of confusion matrices of a-20 and a-06. These show the confusion within the works of the selected author.

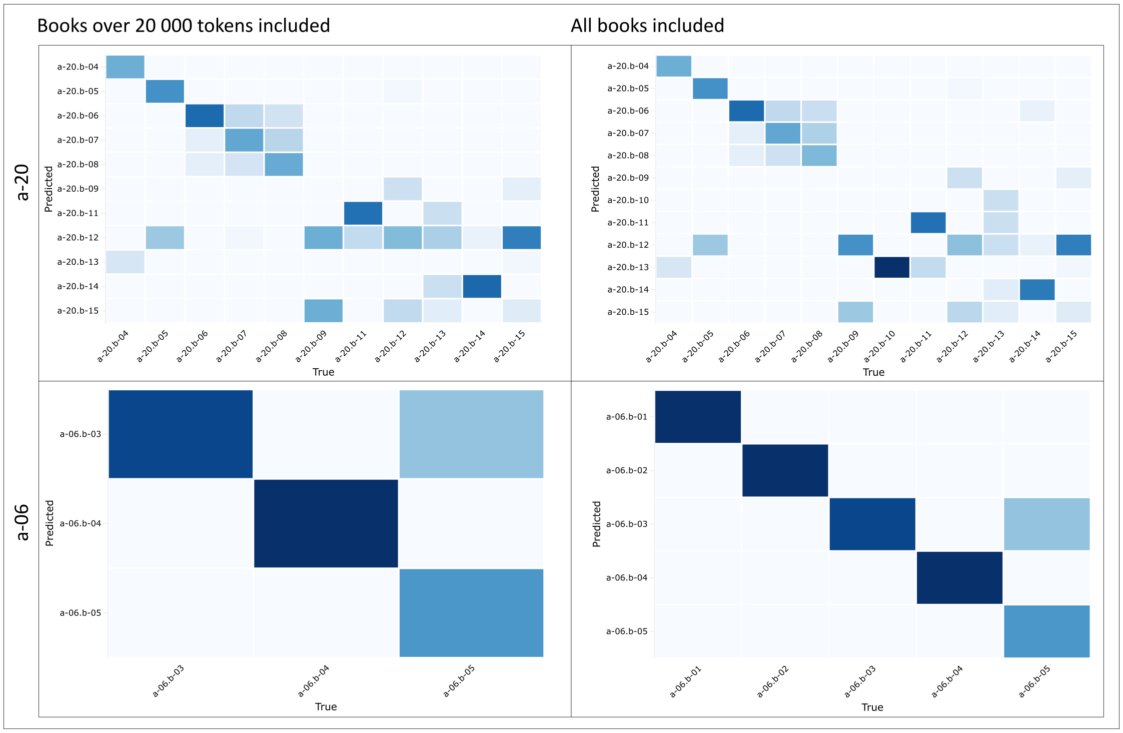

Figure 2. Heatmaps of intra-author classification for a-20 and a-06; created using Flourish (https://flourish.studio/, accessed 5 April 2023)

58 These experiments show that, albeit delexicalised, most books tend to be recognised as themselves, meaning that most authors do write individual books differently, even when the differences in the lexicon are suppressed. At the same time, the results of Step 2 conclusively demonstrate that most authors are still recognisable when tested on an unseen book.

59 Some interesting phenomena revealed by this experiment may be noted. The results show that accuracy scores are not strongly dependent on the books' lengths (the correlation coefficient is ca. 0.27). However, the books under 20,000 tokens (17 books in our dataset) lead to poorer results on average. The accuracy score was only 0.54 (21 of 39 test segments attributed correctly), and six of these books were not recognised at all. In comparison, the accuracy scores of the full model were 0.65 (all books) and 0.66 (only books over 20,000 tokens). At the same time, even some of the longer books performed very badly. For example, of 20 test segments from a-23.b-04, only one was attributed correctly. However, this book scored very well (0.9) in attribution to the correct author (including the one correctly attributed segment) in both experiments. The explanation for this may become clear when the nature of the book is considered — it is a collection of short stories.

60 From this, one may assume that collections of short stories are good candidates for high levels of intra-author confusion. However, some examples contradict this assumption. For example, a-15.b-02, 03, 04, and 06 are collections of short stories by K. Čapek. In contrast to a-23.b-04, these perform quite well in the correct book attribution (0.56 for the experiment with all books, 0.61 for the experiment with over 20,000 token books). But the performance for correct author attribution was not especially high (0.71 in both experiments). Further exploration of features and their weights may help us to understand these differences better.

61 Another interesting example is a high level of confusion among books a-20.b-06, 07, and 08 (see the image above). These three books form a trilogy together (*Francouzská revoluce*), and confusion was, therefore, to be expected.

62 Comparison of steps 2 and 3 may help us to further explore the deviations in the performances of different authors. The following table shows correlations of authors' accuracy scores from Step 2 (using all books irrespective of their token length) to different data obtained from Step 3 related to the authors' test books: (a) proportion of segments attributed to the same book, (b) proportion of segments attributed to other books by the same author, (c) proportion of segments attributed to other authors, and (d) proportion of segments attributed to the correct author (both correct and incorrect

books).

	Set 1	Set 2	Set 3	Set 4	Set 5
Attribution to Correct Book	0.183381	-0.23135	-0.51682	-0.26735	0.523796
To Correct Author // Incorrect Book	0.486218	0.443661	0.441649	0.416486	0.413774
To Incorrect Author	-0.87451	-0.15772	0.286852	-0.27283	-0.88701
To Correct Author	0.872354	0.154229	-0.293	0.270222	0.885652

Table 7. Correlation coefficients of authors' accuracy scores (Step 2) to different results of experiments from Step 3 of the test books attribution.

The correlations shown in this table seem to support our initial assumption that misclassification is a good proxy for similarity. The following table presents the data for a detailed exploration of these correlations. Variations among individual authors are still significant. Other criteria must always be considered.

	Proportion of the Test Book Segments Attributed to the Correct Author but an Incorrect Book (Step 3)					Author's Performance (Step 2)				
	Set 1	**Set 2**	**Set 3**	**Set 4**	**Set 5**	**Set 1**	**Set 2**	**Set 3**	**Set 4**	**Set 5**
a-01	0.54	0.03	0.17	0.00	0.17	0.98	0.91	1.00	0.97	0.82
a-02	0.00	0.50	0.33	0.33	0.05	1.00	0.53	0.94	0.88	0.83
a-03	1.00	0.06	0.00	0.00	0.27	1.00	1.00	0.93	0.20	0.99
a-04	0.38	0.14	0.89	0.25	0.47	0.83	0.90	0.91	0.67	0.97
a-05	0.00	0.16	0.14	0.50	0.38	0.95	0.81	0.85	0.90	0.94
a-06	0.12	0.44	0.08	0.67	0.36	0.64	0.96	0.83	0.78	0.83
a-07	0.80	0.11	0.02	0.50	0.10	0.98	0.91	0.70	0.99	0.99
a-08	0.15	0.10	0.47	0.25	0.32	0.00	0.13	0.67	0.10	0.00
a-09	0.00	0.00	0.12	0.25	0.40	0.45	0.42	0.91	0.24	0.80
a-10	0.25	0,33	0,33	0,00	0,09	0,97	0,79	1,00	0,53	0,93
a-11	0.17	0.21	0.18	0.23	0.50	0.86	0.78	1.00	0.74	0.93
a-12	0.43	0.25	0.00	0.00	0.00	0.66	0.86	0.53	0.89	0.32
a-13	0.31	0.38	1.00	0.17	0.64	0.62	0.89	0.90	1.00	0.92
a-14	0.16	0.21	0.08	0.00	0.00	1.00	1.00	1.00	0.96	1.00
a-15	0.00	0.00	0.11	0.00	0.00	0.24	0.96	0.84	0.83	0.84
a-16	0.00	0.00	0.33	0.00	0.00	0.90	0.96	0.99	0.93	0.95
a-17	0.23	0.26	0.00	0.47	0.51	1.00	0.86	0.96	0.75	1.00
a-18	0.00	0.40	0.00	0.00	0.00	0.98	0.98	0.96	1.00	0.95
a-19	0.17	0.00	0.06	1.00	0.33	0.86	1.00	0.99	0.89	0.97
a-20	0.11	0.05	0.06	0.00	0.14	1.00	0.84	0.98	1.00	0.99
a-21	0.43	0.63	0.53	0.06	0.12	1.00	0.93	0.55	0.65	0.98
a-22	1.00	0.08	0.33	0.06	0.17	0.79	1.00	0.94	0.56	0.88
a-23	0.13	0.21	0.69	0.67	0.17	1.00	0.98	1.00	1.00	0.98

Table 8. Correct author // incorrect book attribution of test books (Step 3) and authors' performance.

Conclusion

We do not claim to have studied the issue of book versus authorial style exhaustively. What we have done is build a

64 pipeline to show that this issue is worth taking into account in developing and evaluating authorship classification systems. It is clear that the performance drops significantly when an unseen book is used for testing instead of unseen segments of books seen during the training process. On the one hand, this is not a catastrophic problem, as performances only drop by 10-20% on average. On the other hand, however, this makes an important difference for applications since it results in a significant rise in the number of errors.

65 The results of our experiments focused on attributing individual books instead of authors have revealed that models trained and tested on the segments from the same book perform well despite a relatively high level of delexicalisation. These experiments have also shown that misclassification of a book but correct classification of an author is a good proxy for similarity in author style. Thus, we may recommend such an experiment in the classifiers' development stage. Further research might also focus on measuring this effect across languages.

Acknowledgements

66 Realised with the support of *Institutional Research* of the National Library of the Czech Republic, funded by the Ministry of Culture of the Czech Republic as part of the framework of *Long-Term Conception Developement of Scientific Organisation*.

Appendix

Author	Title	Genre	Book ID in Dataset	Token Count	Set in Which Used as Test Book (Experiment Step 2)
A. Stašek	*Nedokončený obraz*	Prose	a-01.b-01	91,746	Set 5
A. Stašek	*Otřelá kolečka*	Prose	a-01.b-02	83,978	
A. Stašek	*Vzpomínky*	Prose	a-01.b-03	155,266	
A. Stašek	*Bohatství*	Prose	a-01.b-04	54,474	
A. Stašek	*Bratři*	Prose	a-01.b-05	66,637	Set 3
A. Stašek	*Blouznivci našich hor*	Prose	a-01.b-07	141,011	Set 2
A. Stašek	*O ševci Matoušovi a jeho přátelích*	Prose	a-01.b-08	83,005	
A. Stašek	*Na rozhraní*	Prose	a-01.b-09	106,018	
A. Stašek	*V temných vírech (1)*	Prose	a-01.b-11	89,013	
A. Stašek	*V temných vírech (3)*	Prose	a-01.b-12	92,030	Set 4
A. Stašek	*Stíny minulosti*	Prose	a-01.b-13	81,008	Set 1
A. Stašek Full Tokens Count: 1,044,186					
J. Neruda	*Arabesky*	Prose	a-02.b-01	69,981	Set 2
J. Neruda	*Trhani*	Prose	a-02.b-02	14,772	Set 1
J. Neruda	*Menší cesty*	Prose	a-02.b-03	76,567	Set 3
J. Neruda	*Povídky malostranské*	Prose	a-02.b-04	89,079	Set 4
J. Neruda	*Studie, krátké a kratší*	Prose	a-02.b-05	122,183	Set 5
J. Neruda Full Tokens Count: 364,582					
J. Arbes	*Ethiopská lilie*	Prose	a-03.b-01	79,873	
J. Arbes	*Kandidáti existence*	Prose	a-03.b-02	81,821	
J. Arbes	*Poslední dnové lidstva*	Prose	a-03.b-03	88,181	
J. Arbes	*Persekuce lidu českého v letech 1869-1873*	Prose	a-03.b-04	163,125	Set 4

J. Arbes	*Svatý Xaverius*	Prose	a-03.b-05	28,370	
J. Arbes	*Elegie a idyly*	Prose	a-03.b-06	159,003	
J. Arbes	*Moderní upíři*	Prose	a-03.b-09	93,009	
J. Arbes	*Anděl míru*	Prose	a-03.b-10	106,028	
J. Arbes	*Sivooký démon*	Prose	a-03.b-11	89,031	Set 3
J. Arbes	*Štrajchpudlíci*	Prose	a-03.b-12	125,003	Set 5
J. Arbes	*Akrobati*	Prose	a-03.b-13	40,001	Set 2
J. Arbes	*Divotvorci tónů*	Prose	a-03.b-15	73,023	Set 1
J. Arbes	*Z víru života*	Prose	a-03.b-16	71,002	
J. Arbes Full Tokens Count: 1,197,470					
K. Klostermann	*Ze světa lesních samot*	Prose	a-04.b-01	87,234	Set 3
K. Klostermann	*Za štěstím*	Prose	a-04.b-02	95,745	Set 1
K. Klostermann	*Domek v Polední ulici*	Prose	a-04.b-03	100,419	Set 2
K. Klostermann	*Vypovězen*	Prose	a-04.b-04	70,129	Set 5
K. Klostermann	*Kulturní naléhavost*	Prose	a-04.b-05	18,382	Set 4
K. Klostermann Full Tokens Count: 371,909					
F. X. Šalda	*Boje o zítřek*	Prose	a-05.b-01	63,141	Set 1
F. X. Šalda	*Moderní literatura česká*	Prose	a-05.b-02	16,843	Set 5
F. X. Šalda	*Duše a dílo*	Prose	a-05.b-03	82,283	Set 3
F. X. Šalda	*Umění a náboženství*	Prose	a-05.b-04	10,141	Set 4
F. X. Šalda	*Juvenilie: stati, články a recense z let 1891-1899 (1)*	Prose	a-05.b-05	112,978	Set 2
F. X. Šalda Full Tokens Count: 285,386					
T. G. Masaryk	*Blaise Pascal, jeho život a filosofie*	Prose	a-06.b-01	11,662	Set 3
T. G. Masaryk	*O studiu děl básnických*	Prose	a-06.b-02	8,786	Set 4
T. G. Masaryk	*Česká otázka: snahy a tužby národního obrození*	Prose	a-06.b-03	63,168	Set 5
T. G. Masaryk	*Otázka sociální: základy marxismu sociologické a filosofické*	Prose	a-06.b-04	190,279	Set 1
T. G. Masaryk	*Jan Hus: naše obrození a naše reformace*	Prose	a-06.b-05	25,635	Set 2
T. G. Masaryk Full Tokens Count: 299,530					
A. Jirásek	*Na Chlumku*	Prose	a-07.b-02	8,016	
A. Jirásek	*Na dvoře vévodském*	Prose	a-07.b-04	81,005	
A. Jirásek	*Psohlavci*	Prose	a-07.b-05	88,007	
A. Jirásek	*Zahořanský hon a jiné povídky*	Prose	a-07.b-06	75,002	Set 2
A. Jirásek	*Skály*	Prose	a-07.b-07	90,021	
A. Jirásek	*Temno*	Prose	a-07.b-08	215,002	Set 1
A. Jirásek	*Bratrstvo (1): Bitva u Lučence*	Prose	a-07.b-09	146,023	Set 5
A. Jirásek	*Bratrstvo (2): Mária*	Prose	a-07.b-10	158,003	
A. Jirásek	*Bratrstvo (3): Žebráci*	Prose	a-07.b-11	180,009	Set 4
A. Jirásek	*F.L. Věk*	Prose	a-07.b-12	152,028	

A. Jirásek	*Maryla*	Prose	a-07.b-13	53,035	
A. Jirásek	*Husitský král (2)*	Prose	a-07.b-13	115,006	
A. Jirásek	*Lucerna*	Drama	a-07.b-14	27,001	Set 3
A. Jirásek	*Mezi proudy (1)*	Prose	a-07.b-16	124,009	
A. Jirásek Full Tokens Count: 1,512,167					
Č. Slepánek	*Srbsko od prvého povstání 1804 do dnešní doby*	Prose	a-08.b-01	110,022	Set 2
Č. Slepánek	*Črty z Ruska a odjinud*	Prose	a-08.b-02	40,032	Set 4
Č. Slepánek	*Svědomí Lidových novin, čili, Jak bylo po léta v českém tisku štváno lživě proti mně*	Prose	a-08.b-03	6,004	Set, Set 5
Č. Slepánek	*Dělnické hnutí v Rusku*	Prose	a-08.b-04	18,057	Set 3
Č. Slepánek Full Tokens Count: 174,115					
E. Krásnohorská	*Svéhlavička*	Prose	a-09.b-01	74,030	Set 4
E. Krásnohorská	*Celínka*	Prose	a-09.b-02	90,003	Set 5
E. Krásnohorská	*Pohádky Elišky Krásnohorské*	Prose	a-09.b-03	40,004	Set 1
E. Krásnohorská	*Srdcem i skutkem*	Prose	a-09.b-04	24,032	
E. Krásnohorská	*Do proudu žití*	Prose	a-09.b-06	47,013	Set 3
E. Krásnohorská	*Medvěd a víla*	Drama	a-09.b-08	12,002	Set 2
E. Krásnohorská	*Čertova stěna*	Drama	a-09.b-10	14,003	
E. Krásnohorská	*Trojí máj*	Prose	a-09.b-11	73,017	
E. Krásnohorská Full Tokens Count: 374,104					
F. Herites	*Amanita*	Prose	a-10.b-01	73,015	Set 2
F. Herites	*Tajemství strýce Josefa*	Prose	a-10.b-02	52,010	
F. Herites	*Maloměstské humoresky*	Prose	a-10.b-03	69,021	Set 3
F. Herites	*Tři cesty*	Prose	a-10.b-04	28,010	
F. Herites	*Bez chleba*	Prose	a-10.b-06	92,013	
F. Herites	*Všední zjevy*	Prose	a-10.b-07	99,011	Set 1
F. Herites	*Bůh v lidu*	Prose	a-10.b-09	11,022	
F. Herites	*Vodňanské vzpomínky*	Prose	a-10.b-10	19,009	Set 4
F. Herites	*Sebrané spisy Fr. Heritesa*	Prose	a-10.b-11	71,020	Set 5
F. Herites Full Tokens Count: 514,131					
I. Olbracht	*Nikola Šuhaj loupežník*	Prose	a-11.b-01	67,028	
I. Olbracht	*Anna proletářka*	Prose	a-11.b-02	81,016	Set 1
I. Olbracht	*Karavany v noci*	Prose	a-11.b-03	99,007	
I. Olbracht	*Žalář nejtemnější*	Prose	a-11.b-04	41,002	Set 2
I. Olbracht	*Dobyvatel*	Prose	a-11.b-05	193,020	
I. Olbracht	*O smutných očích Hany Karadžičové*	Prose	a-11.b-06	46,004	Set 3
I. Olbracht	*O zlých samotářích*	Prose	a-11.b-07	117,007	Set 4

I. Olbracht	*Golet v údolí*	Prose	a-11.b-08	71,009	Set 5
I. Olbracht Full Tokens Count: 715,093					
J. Vrchlický	*Povídky ironické a sentimentální*	Prose	a-12.b-01	25,041	Set 5
J. Vrchlický	*Barevné střepy*	Prose	a-12.b-03	26,001	
J. Vrchlický	*Nové barevné střepy*	Prose	a-12.b-05	35,002	Set 1
J. Vrchlický	*Loutky*	Prose	a-12.b-06	84,012	
J. Vrchlický	*Noc na Karlštejně*	Drama	a-12.b-07	21,002	Set 2
J. Vrchlický	*Drahomíra*	Drama	a-12.b-08	15,010	Set 3
J. Vrchlický	*Knížata*	Drama	a-12.b-09	35,043	Set 4
J. Vrchlický Full Tokens Count: 241,111					
J.S. Machar	*Nemocnice*	Prose	a-13.b-01	34,020	Set 1
J.S. Machar	*Pod sluncem italským*	Prose	a-13.b-01	57,027	
J.S. Machar	*Třicet roků*	Prose	a-13.b-03	60,014	Set 5
J.S. Machar	*Vídeň*	Prose	a-13.b-04	68,009	
J.S. Machar	*Řím*	Prose	a-13.b-05	69,005	Set 4
J.S. Machar	*Vzpomíná se…*	Prose	a-13.b-06	70,002	Set 2
J.S. Machar	*Kriminál*	Prose	a-13.b-07	59,003	Set 3
J.S. Machar Full Tokens Count: 417,080					
J. Zeyer	*Ondřej Černyšev*	Prose	a-14.b-01	91,005	
J. Zeyer	*Román o věrném přátelství Amise a Amila*	Prose	a-14.b-02	91,036	
J. Zeyer	*Báje Šošany*	Prose	a-14.b-03	43,010	Set 1
J. Zeyer	*Fantastické povídky*	Prose	a-14.b-04	82,017	
J. Zeyer	*Dobrodružství Madrány*	Prose	a-14.b-05	57,017	Set 5
J. Zeyer	*Gompači a Komurasaki*	Prose	a-14.b-06	38,011	
J. Zeyer	*Rokoko: Sestra Paskalina*	Prose	a-14.b-07	30,001	Set 2
J. Zeyer	*Jan Maria Plojhar*	Prose	a-14.b-08	115,022	Set 4
J. Zeyer	*Stratonika a jiné povídky*	Prose	a-14.b-09	91,026	Set 3
J. Zeyer	*Maeldunova výprava a jiné povídky*	Prose	a-14.b-10	34,046	
J. Zeyer	*Tři legendy o krucifixu*	Prose	a-14.b-11	59,016	
J. Zeyer Full Tokens Count: 731,207					
K. Čapek	*Válka s mloky*	Prose	a-15.b-01	83,021	Set 3
K. Čapek	*Nůše pohádek (3)*	Prose	a-15.b-02	42,020	Set 1
K. Čapek	*Povídky z jedné kapsy*	Prose	a-15.b-03	61,027	
K. Čapek	*Povídky z druhé kapsy*	Prose	a-15.b-04	52,019	
K. Čapek	*Věc Makropulos*	Drama	a-15.b-05	22,007	
K. Čapek	*Devatero pohádek*	Prose	a-15.b-06	56,004	
K. Čapek	*Ze života hmyzu*	Drama	a-15.b-07	22,004	
K. Čapek	*Měl jsem psa a kočku*	Prose	a-15.b-08	25,021	
K. Čapek	*Matka*	Drama	a-15.b-09	24,005	Set 2
K. Čapek	*Zahradníkův rok*	Prose	a-15.b-10	25,007	
K. Čapek	*Povětroň*	Prose	a-15.b-11	52,003	
K. Čapek	*Jak se co dělá*	Prose	a-15.b-12	34,004	
K. Čapek	*Loupežník*	Drama	a-15.b-13	23,003	Set 4

K. Čapek	*Cesta na sever*	Prose	a-15.b-14	33,003	
K. Čapek	*Hovory s T.G. Masarykem*	Prose	a-15.b-15	24,013	
K. Čapek	*Továrna na Absolutno, Krakatit*	Prose	a-15.b-16	147,012	
K. Čapek	*Bílá nemoc*	Drama	a-15.b-17	25,003	Set 5
K. Čapek	*Boží muka*	Prose	a-15.b-18	35,022	
K. Čapek Full Tokens Count: 785,198					
K. Nový	*Plamen a vítr*	Prose	a-16.b-01	174,008	
K. Nový	*Železný kruh*	Prose	a-16.b-02	300,021	Set 2
K. Nový	*Peníze*	Prose	a-16.b-03	77,003	Set 3
K. Nový	*Chceme žít*	Prose	a-16.b-04	58,001	
K. Nový	*Na rozcestí*	Prose	a-16.b-05	126,002	Set 5
K. Nový	*Atentát*	Prose	a-16.b-06	113,009	
K. Nový	*Rytíři a lapkové*	Prose	a-16.b-07	137,001	Set 1
K. Nový	*Balada o českém vojáku*	Prose	a-16.b-08	47,054	
K. Nový	*Rybaříci na Modré zátoce*	Prose	a-16.b-09	23,001	
K. Nový	*Potulný lovec*	Prose	a-16.b-10	44,003	Set 4
K. Nový Full Tokens Count: 1,099,103					
K. Sabina	*Synové světla*	Prose	a-17.b-01	230,005	Set 3
K. Sabina	*Hrobník*	Prose	a-17.b-02	27,001	Set 1
K. Sabina	*Morana čili Svět a jeho nicoty*	Prose	a-17.b-03	144,003	Set 2
K. Sabina	*Oživené hroby*	Prose	a-17.b-04	86,020	
K. Sabina	*Černá růže*	Drama	a-17.b-05	20,002	Set 4
K. Sabina	*Blouznění*	Prose	a-17.b-07	107,001	Set 5
K. Sabina Full Tokens Count: 614,032					
K.V. Rais	*Zapadlí vlastenci*	Prose	a-18.b-01	125,026	
K.V. Rais	*Maloměstské humorky*	Prose	a-18.b-02	128,004	Set 5
K.V. Rais	*Kalibův zločin*	Prose	a-18.b-03	65,028	
K.V. Rais	*Paničkou: obraz z podhoří*	Prose	a-18.b-04	60,008	Set 4
K.V. Rais	*Povídky o českých umělcích*	Prose	a-18.b-05	22,004	
K.V. Rais	*Povídky ze starých hradů*	Prose	a-18.b-07	32,012	
K.V. Rais	*Výminkáři*	Prose	a-18.b-09	48,001	
K.V. Rais	*Stehle: podhorský obraz*	Prose	a-18.b-10	124,023	Set 3
K.V. Rais	*Z rodné chaloupky*	Prose	a-18.b-11	23,008	
K.V. Rais	*Skleník*	Prose	a-18.b-12	33,004	
K.V. Rais	*Pantáta Bezoušek*	Prose	a-18.b-13	88,006	Set 1
K.V. Rais	*Ze srdce k srdcím*	Prose	a-18.b-14	22,002	
K.V. Rais	*Horské kořeny*	Prose	a-18.b-15	49,019	Set 2
K.V. Rais Full Tokens Count: 819,145					
K. Světlá	*Černý Petříček*	Prose	a-19.b-01	35,025	
K. Světlá	*Poslední poustevnice*	Prose	a-19.b-02	52,001	
K. Světlá	*Z let probuzení*	Prose	a-19.b-03	70,037	Set 4
K. Světlá	*Na úsvitě*	Prose	a-19.b-04	108,002	
K. Světlá	*Kantůrčice*	Prose	a-19.b-05	65,001	Set 5
K. Světlá	*O krejčíkově Anežce*	Prose	a-19.b-06	21,011	Set 1

K. Světlá	*Časové ohlasy*	Prose	a-19.b-07	72,044	Set 3
K. Světlá	*Kříž u potoka*	Prose	a-19.b-08	102,025	
K. Světlá	*Vesnický román*	Prose	a-19.b-09	77,015	
K. Světlá	*Frantina*	Prose	a-19.b-10	65,001	
K. Světlá	*Nemodlenec*	Prose	a-19.b-11	98,035	Set 2
K. Světlá Full Tokens Count: 765,197					
S.K. Neumann	*Československá cesta*	Prose	a-20.b-04	32,009	
S.K. Neumann	*Vzpomínky (1)*	Prose	a-20.b-05	40,006	Set 1
S.K. Neumann	*Francouzská revoluce (1)*	Prose	a-20.b-06	158,001	
S.K. Neumann	*Francouzská revoluce (2)*	Prose	a-20.b-07	171,012	Set 5
S.K. Neumann	*Francouzská revoluce (3)*	Prose	a-20.b-08	157,013	
S.K. Neumann	*Ať žije život*	Prose	a-20.b-09	42,022	
S.K. Neumann	*Jelec*	Prose	a-20.b-10	11,008	
S.K. Neumann	*Enciány s Popa Ivana*	Prose	a-20.b-11	24,012	Set 4
S.K. Neumann	*O umění*	Prose	a-20.b-12	217,009	
S.K. Neumann	*Paměti a drobné prózy*	Prose	a-20.b-13	47,018	Set 3
S.K. Neumann	*Zlatý oblak*	Prose	a-20.b-14	70,018	Set 2
S.K. Neumann	*Konfese a konfrontace (2)*	Prose	a-20.b-15	168,005	
S.K. Neumann Full Tokens Count: 1,137,133					
V. Hálek	*Na vejminku*	Prose	a-21.b-01	46,020	
V. Hálek	*Pod pustým kopcem*	Prose	a-21.b-03	58,023	Set 5
V. Hálek	*Mejrima a Husejn*	Poetry	a-21.b-04	17,009	Set 4
V. Hálek	*Král Rudolf*	Drama	a-21.b-06	25,012	
V. Hálek	*Komediant*	Prose	a-21.b-08	87,019	Set 2
V. Hálek	*Na statku a v chaloupce*	Prose	a-21.b-09	38,004	Set 1
V. Hálek	*Kresby křídou i tuší*	Prose	a-21.b-10	146,014	
V. Hálek	*Povídky I*	Prose	a-21.b-11	116,005	
V. Hálek	*Fejetony*	Prose	a-21.b-12	170,015	Set 3
V. Hálek Full Tokens Count: 703,121					
V. Vančura	*Obrazy z dějin národa českého*	Prose	a-22.b-01	141,011	
V. Vančura	*Kubula a Kuba Kubikula*	Prose	a-22.b-02	18,016	Set 4
V. Vančura	*Pole orná a válečná*	Prose	a-22.b-03	46,002	
V. Vančura	*Amazonský proud; Dlouhý, Široký, Bystrozraký*	Prose	a-22.b-04	38,002	Set 1
V. Vančura	*Pekař Jan Marhoul*	Prose	a-22.b-05	34,015	Set 2
V. Vančura	*Poslední soud*	Prose	a-22.b-06	37,004	

V. Vančura	*Luk královny Dorotky*	Prose	a-22.b-07	33,001	Set 3
V. Vančura	*Tři řeky*	Prose	a-22.b-08	93,014	
V. Vančura	*Rozmarné léto*	Prose	a-22.b-10	23,011	
V. Vančura	*Markéta Lazarová*	Prose	a-22.b-11	46,008	
V. Vančura	*Rodina Horvatova*	Prose	a-22.b-12	109,005	Set 5
V. Vančura Full Tokens Count: 618,089					
Z. Winter	*Nezbedný bakalář a jiné rakovnické obrázky*	Prose	a-23.b-01	115,003	Set 4
Z. Winter	*Ze staré Prahy*	Prose	a-23.b-02	62,005	Set 5
Z. Winter	*Krátký jeho svět a jiné pražské obrázky*	Prose	a-23.b-04	102,009	
Z. Winter	*Staré listy*	Prose	a-23.b-05	66,007	Set 1
Z. Winter	*Rozina sebranec*	Prose	a-23.b-06	64,019	Set 3
Z. Winter	*Bouře a přeháňka*	Prose	a-23.b-07	69,001	
Z. Winter	*Panečnice*	Prose	a-23.b-08	28,025	
Z. Winter	*Mistr Kampanus*	Prose	a-23.b-09	177,039	Set 2
Z. Winter Full Tokens Count: 683,108					

Table 9. List of works in dataset.

Notes

[1] See, for example, [Savoy 2020], [Swain, Mishra, and Sindhu 2017], [Grzybek 2014], [Grieve 2005], and [Holmes 1998] for general studies following the development of the area. For literary oriented studies that cover a more or less random selection of works we have consulted during our research, see [Zhao and Zobel 2007], [Kusakci 2012], [Segarra, Eisen, and Ribeiro 2013], [Ramezani, Sheydaei, and Kahani 2013], [Pinho, Pratas, and Ferreira 2016], [Nutanong et al.], [Marinho, Hirst, and Amancio 2016], [Benotto 2021], and [Gorman 2022]. We also organized a workshop, “Authorial style, its analysis, and limits of automatic recognition”, at the National Library of the Czech Republic in 2022, which brought together research approaching the topic from diverse perspectives, demonstrating the rich and complex problematics of authorial style detection. See https://digilab.nkp.cz/?page_id=55 (accessed 5 April 2023).

[2] Stylo: Stylometric Multivariate Analyses, available at https://cran.r-project.org/package=stylo (accessed 5 April 2023).

[3] Compare also with [Luyckx and Daelemans 2011], who focus on the effect of author set size and data size in authorship attribution, taking into consideration a variety of genres and topics. Luyckx and Daelemans' use cases focus on much shorter texts than this article does or Gorman, thus posing a different issue.

[4] We have used the digital collections of the National Library of the Czech Republic (https://ndk.cz/, accessed April 5, 2023) as the source of our data. Unfortunately, these data are not publicly accessible, which creates issues regarding the repeatability of our experiments.

[5] The raw data consisted of individual pages as .txt files with inconsistent encoding. Firstly, the encoding was unified to UTF-8. From these files we attempted to remove non-content data such as headers, page numbers, footnotes, etc. This process was automatized and therefore may include some imperfections. After this initial cleaning, we merged the individual pages into a single .txt file per book.

[6] Available at: http://lindat.mff.cuni.cz/services/korektor/, accessed 5 April 2023.

[7] In machine learning experiments, the development set is used to evaluate different hyperparameter settings (such as regularization strength or internal dimension of the model) and models in order to select the best model and its setting. Once all these choices are fixed, the selected model is trained on a combination of the training and development sets, and the test set is used to estimate the expected system performance on unseen data. If one used the test set rather than the development set for hyperparameter optimization, the final evaluation result would be artificially inflated by information leakage from the test set into the hyperparameter design — hence the use of a development set.

[8] Available at: https://lindat.mff.cuni.cz/services/udpipe/api-reference.php, accessed 5 April 2023; see [Straka 2018].

[9] Autosemantic words, as recognized by UDPipe, are: nouns (NOUN), proper nouns (PROPN), adjectives (ADJ), verbs (VERB), adverbs (ADV), and numbers (NUM).

[10] See http://universaldependencies.org/docs/u/pos/index.html, accessed April 5, 2023.

[11] These are the books designated as b-01 to b-05 for each of the authors in the list of works in the appendix.

[12] Codes r-01, r-02, and r-03 were used in preparation for further delexicalisation; therefore, we start with r-04.

[13] Aside from the word form for non-delexicalised baselines, we have used the lemma, the full morphological tag according to the Universal Dependencies specification, and at the coarsest level of granuality, the universal part-of-speech tag. See http://universaldependencies.org/docs/u/pos/index.html, accessed April 5, 2023.

[14] The types of named entities are persons (first names and surnames), locations, organizations (including brands), and miscellaneous named entities such as religions, sports leagues, and wars. For a detailed list, see https://www.cnts.ua.ac.be/conll2003/ner/annotation.txt, accessed September 15, 2023.

[15] There is clearly space for improving classification accuracy here; the features in [Gorman 2022] reported little such decrease in a comparable experiment. While 42/42.3% accuracy with segments of approximately 50 tokens is still above the 23-class baseline (4.35%), in order to provide useful results outside of long-form texts, the classification pipeline would need significant improvement. Again, we emphasize we are not trying to reach the highest possible accuracy. Rather, we use classification experiments to illustrate variation within an author's style.

[16] For the effect of author set size and data size in authorship attribution, see [Luyckx and Daelemans 2011].

[17] Specifically, the 11 test books where we compared the 5-fold and leave-one-out accuracies were: a-03 Set 4, with a change from 0.20 to 0.276 (+7.6%); a-15 Set 1: 0.24 → 0.214 (-2.6%); a-02 Set 2: 0.53→0.516 (-1.4%); a-07 Set 3: 0.7→0.63 (-7.0%); a-17 Set 4 (drama): 0.75→0.75 (0.0%); a-08 Set 3: 0.67→0.667% (-0.3%); a-09 Set 3: 0.91→0.85 (-6.0%); a-13 Set 4: 1.0→1.0 (0.0%); a-22 Set 2: 1.0→1.0 (0.0%); a-11 Set 3: 1.0→1.0 (0.0%); a-21 Set 4 (poetry): 0.65→0.706 (+5.6%). The average difference when discarding the three books with perfect accuracies was that in the leave-one-out setting, classification was 0.3% worse. Without these three books taken into account (because they may be so easy to classify that even a very flawed methodology pipeline would obtain perfect accuracy), leave-one-out classification performed 0.5% worse than the five-fold setting. (We give the books here as author-set pairs rather than author-books, so that their “outlier-ness” is easy to find in the tables in this section. To find which book these are, refer to Appendix: List of Works in the Dataset.)

[18] E. Krásnohorská (a-09) in Set 2; J. Vrchlický (a-12) in sets 2, 3, 4; K. Čapek (a-15) in sets 2, 4, 5; K. Sabina (a-17) in Set 4; see appendix.

[19] V. Hálek (a-21) in Set 4; see appendix.

Works Cited

Benotto 2021 Benotto, G. (2021) “Can an author style be unveiled through word distribution?”, *Digital Humanities Quarterly*, 15(1). http://digitalhumanities.org:8081/dhq/vol/15/1/000539/000539.html

Gorman 2022 Gorman, R. (2022) “Universal dependencies and author attribution of short texts with syntax alone”, *Digital Humanities Quarterly*, 16(2). http://digitalhumanities.org:8081/dhq/vol/16/2/000606/000606.html

Grieve 2005 Grieve, J. (2005) *Quantitative authorship attribution: A history and an evaluation of techniques*. MA thesis. Simon Fraser University. Available at: https://summit.sfu.ca/item/8840.

Grzybek 2014 Grzybek, P. (2014) “The emergence of stylometry: Prolegomena to the history of term and concept”, in Kroó, K. and Torop, P. (eds.) *Text within text: Culture within Culture*. Paris: L’Harmattan, pp. 58–75.

Holmes 1998 Holmes, D.I. (1998) “The evolution of stylometry in humanities scholarship”, *Literary and Linguistic Computing*, 13(3), pp. 111–117.

Kusakci 2012 Kusakci, A.O. (2012) “Authorship attribution using committee machines with k-nearest neighbors rated voting”, *Proceedings of the 11th symposium on neural network applications in electrical engineering, IEEE, 2012*. Belgrade, Serbia, 20–22 September. pp. 161–166. Available at: https://ieeexplore.ieee.org/document/6419997.

Luyckx 2011 Luyckx, K. (2011) *Scalability issues in authorship attribution*. Brussels, Belgium: University Press Antwerp.

Luyckx and Daelemans 2011 Luyckx, K., and Daelemans, W. (2011) “The effect of author set size and data size in authorship attribution”, *Literary and Linguistic Computing*, 26(1), pp. 35–55.

Marinho, Hirst, and Amancio 2016 Marino, V.Q., Hirst, G., and Amancio, D.R. “Authorship attribution via network motifs identification”, *Proceedings of the 5th Brazilian conference of intelligent systems, IEEE, 2016*. Recife, Brazil, 9–12 October. pp. 355–360. https://doi.org/10.48550/arXiv.1607.06961.

Mosteller and Wallace 1964 Mosteller, F., and Wallace, D. (1964) *Inference and disputed authorship: The Federalist*. Reading, MA: Addison-Wesley.

Nivre 2015 Nivre, J. (2015) “Towards a universal grammar for natural language processing”, *Proceedings of the international conference on intelligent text processing and computational linguistics, CICLing, 2016*. Konya, Turkey, 3–9 April. https://doi.org/10.1007/978-3-319-18111-0_1.

Nutanong et al. Nutanong, S. et al. “A scalable framework for stylometric analysis query processing”, *Proceedings of the 16th international conference on data mining, IEEE, 2016*. Barcelona, Spain, 12–15 December. pp. 1125–1130. https://doi.org/10.1109/ICDM.2016.0147.

Pedregosa et al. 2011 Pedregosa, F. et al. (2011) “Scikit-learn: Machine learning in Python”, *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Pinho, Pratas, and Ferreira 2016 Pinho, A.J., Pratas, D., and Ferreira, P.J.S.G. “Authorship attribution using relative compression”, *Proceedings of the data compression conference, IEEE, 2016*. Snowbird, UT, 30 March–1 April. pp. 329–338. https://doi.org/10.1109/DCC.2016.53.

Ramezani, Sheydaei, and Kahani 2013 Ramezani, R., Sheydaei, N., and Kahani, M. (2013) “Evaluating the effects of textual features on authorship attribution accuracy”, *Proceedings of the international econference on computer and knowledge engineering, IEEE, 2016*. Mashhad, Iran, 31 October–1 November. pp. 108–113. https://doi.org/10.1109/ICCKE.2013.6682828

Savoy 2020 Savoy, J. (2020) *Machine learning methods for stylometry: Authorship attribution and author profiling*. New York: Springer Publishing.

Segarra, Eisen, and Ribeiro 2013 Segarra, S., Eisein, M., and Ribeiro, A. (2013) “Authorship attribution using function words adjaceny networks”, *Proceedings of the international conference on acoustics, speech and signal processing, IEEE, 2013* . Vancouver, Canada, 26–31 May. https://doi.org/10.1109/ICASSP.2013.6638728.

Stamatatos 2009 Stamatatos, E. (2009) “A survey of modern authorship attribution methods”, *Journal of the American Society for Information Science and Technology*, 60(3), pp. 538–556. https://doi.org/10.1002/asi.21001.

Straka 2018 Straka, M. (2018) “UDPipe 2.0 prototype at CoNLL 2018 UD shared task”, *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies, ACL, 2018*, pp. 197–207. https://doi.org/10.18653/v1/K18-2020.

Straková, Straka, and Hajič 2019 Straková, J., Straka, M., and Hajič, J. (2019) “Neural architectures for nested NER through linearization”, *Proceedings of the 57th annual meeting of the association for computational linguistics, ACL, 2019*. Florence, Italy, 28 July–2 August. pp. 5326–5331. Available at: https://aclanthology.org/P19-1527.pdf.

Swain, Mishra, and Sindhu 2017 Swain, S., Mishra, G., and Sinhu, C. (2017) “Recent approaches on authorship attribution techniques: An overview”, *Proceedings of the international conference of electronics, commmunication, and aerospace technology, ICECA, 2017*. Coimbatore, India, 20–22 April. pp. 557–566. https://doi.org/10.1109/ICECA.2017.8203599

Tyo, Dhingra, and Lipton 2022 Tyo, J., Dhingra, B., and Lipton, Z.C. (2022) “On the state of the art in authorship attribution and authorship verification”, *arXiv*. https://doi.org/10.48550/arXiv.2209.06869.

Zhao and Zobel 2007 Zhao, Y, and Zobel, J. (2007) “Searching with style: Authorship attribution in classic literature”, *Proceedings of the 30th Australasian computer science conference, ACSC, 2007*. Ballarat, Australia, 30 January–2 February. pp. 59–68. Available at: https://www.researchgate.net/publication/221574042_Searching_With_Style_Authorship_Attribution_in_Classic_Literature.