D. S. Nikolaev [ab]
ORCID: 0000-0002-3034-9794
✉ *dnikolaev@fastmail.com*
M. V. Shumilin [ac]
ORCID: 0000-0002-4348-3909
✉ *mvlshumilin@gmail.com*
[a] *Российская академия народного хозяйства и государственной службы при Президенте РФ (Россия, Москва)*
[b] *Стокгольмский университет (Швеция, Стокгольм)*
[c] *Институт мировой литературы им. А. М. Горького РАН (Россия, Москва)*

Identifying Latin authors through maximum-likelihood Dirichlet inference: A contribution to model-based stylometry

Аннотация. В статье предлагается новый алгоритм для определения авторов латинских прозаических текстов, основанный на Дельте Берроуза и распределении Дирихле. Для демонстрации эффективности алгоритма проводится анализ фрагментов текстов 36 авторов классического и средневекового периода. Наш алгоритм показывает результаты, сопоставимые с результатами, полученными за счет применения Random Forest, одного из самых мощных универсальных классификационных алгоритмов. Преимущество нашего алгоритма заключается в том, что он требует очень мало времени и вычислительных ресурсов для обучения, его легко имплементировать на любом языке программирования общего назначения и его тривиально параллелизовать. Кроме того, поскольку алгоритм основан на эксплицитной модели порождения текста, параметры натренированной модели поддаются интерпретации: точность распределения (сумма его параметров) прямо соответствует стилистической гомогенности текстов соответствующего автора.

Статья подготовлена в рамках выполнения научно-исследовательской работы государственного задания РАНХиГС.

Ключевые слова: стилометрия, латинская литература, распределение Дирихле, Дельта Берроуза, Random Forest, атрибуция текстов, стилистический анализ, машинное обучение

Для цитирования: *Nikolaev D. S., Shumilin M. V.* Identifying Latin authors through maximum-likelihood Dirichlet inference: A contribution to model-based stylometry // Шаги/Steps. Т. 7. № 1. 2021. С. 183–198. https://doi.org/10.22394/2412-9410-2021-7-1-183-198.

Статья поступила в редакцию 9 апреля 2020 г.
Принято к печати 3 августа 2020 г.

D. S. Nikolaev [ab]
ORCID: 0000-0002-3034-9794
✉ *dnikolaev@fastmail.com*
M. V. Shumilin [ac]
ORCID: 0000-0002-4348-3909
✉ *mvlshumilin@gmail.com*
[a] *The Russian Presidential Academy of National Economy and Public Administration (Russia, Moscow)*
[b] *Stockholm University (Sweden, Stockholm)*
[c] *A. M. Gorky Institute of World Literature of the Russian Academy of Sciences (Russia, Moscow)*

Identifying Latin authors through maximum-likelihood Dirichlet inference: A contribution to model-based stylometry

Abstract. The last two decades saw a dramatic increase in the number of papers published on the subject of stylometry, which is often narrowly understood as the task of identification of the author of a particular text fragment based on its stylistic properties. We present a new lightweight algorithm for stylometric identification of authors of Latin prose texts based on Burrows's Delta, computed over relative frequencies of 244 manually selected genre and topic neutral words, and the Dirichlet distribution, whose parameters we estimate using an iterative maximum-likelihood algorithm. In order to demonstrate the effectiveness of the method, we present a case study of 3000-word fragments of texts by 36 classical and medieval authors and show that our method performs on par with Random Forest, a powerful general-purpose classification algorithm. We provide summary statistics of our algorithm's performance together with confusion matrices demonstrating pairwise discriminability of texts by different authors. The advantages of our method are that it is very simple to implement, very quick to train and do inference with, and that it is very interpretable since it is a model-based algorithm: precision of the fitted Dirichlet distributions directly corresponds to the stylistic homogeneity of the texts by different authors. This makes it possible to use the algorithm as a general research tool in Latin stylistics.

The article was written on the basis of the RANEPA state assignment research programme.

Keywords: stylometry, Latin literature, Dirichlet distribution, Burrows's Delta, Random Forest, text attribution, stylistic analysis, machine learning

To cite this article: Nikolaev, D. S., & Shumilin, M. V. (2021). Identifying Latin authors through maximum-likelihood Dirichlet inference: A contribution to model-based stylometry. *Shagi/Steps, 7*(1), 183–198. https://doi.org/10.22394/2412-9410-2021-7-1-183-198.

Received April 9, 2020
Accepted August 3, 2020

1. Introduction

The last two decades saw a dramatic increase in the number of papers published on the subject of stylometry, which is often narrowly understood as the task of identification of the author of a particular text fragment based on its stylistic properties. Stylometry was launched as a computational discipline in the 1960s with the celebrated analysis of the Federalist Papers in [Mosteller, Wallace 1964]; see [Holmes 1998; Holmes, Kardos 2003] for an overview of the early developments.

The statistical-learning boom of the early 21st century, following the seminal early work, such as [Vapnik 1999; Breiman 2001], gave rise to a plethora of new algorithms and approaches to the analysis of textual data. Many of those algorithms have a natural application in stylometry, while others were created specifically for this purpose.

Contemporary approaches can be roughly divided in two groups:

1. Feature-based approaches rely on features extracted from texts to ascertain their authorship. Counts of function words, most-frequent words, or word or character N-grams, and different combinations thereof are usually employed as features. Texts are then clustered or directly compared based on some distance measure, such as Burrows's Delta [Burrows 2002]. Function words have been traditionally considered as good features for stylometric analysis because they are not tied to particular genres and it is hard for authors to deliberately manipulate their frequencies.

2. Model-based approaches aim to directly model the distribution of features in texts by different authors. A text-generating model for an author is created, which makes it possible to directly estimate the posterior probability (in a Bayesian setting) or likelihood (in a frequentist one) that the fragment of interest was composed by this author. Decisions about authorship are then made based on these estimates.

Work on stylometric attribution of Latin texts was mostly done using the discriminative approach. A notable, albeit controversial example, is the attempt by Justin Stover to prove that the Latin fragment preserved in the 13th-century manuscript MS Vat. Reg. lat. 1572 is the long-lost Book 3 of the treaty *De dogmate Platonis* by Apuleius. In his edition of the text [Stover 2015] and several co-authored articles [Stover et al. 2016; Stover, Kestemont 2016a], Stover employed PCA, Burrows's Delta-based Bootstrap-Consensus Trees [Eder et al. 2016], and the Impostor Method [Koppel, Winter 2014] in order to substantiate his claim. Stover and his co-authors also tried to apply a similar methodology to the problem of the authorship of other texts, including the *Corpus Caesarianum* [Kestemont et al. 2016]. Related work was reported in [Kabala 2020], who applied a distance-based classification ("let text segment A be attributed correctly if the next closest text segment in its cor-

pus belongs to the same author class as A") and logistic regression to the problem of the authorship of the twelfth-century Latin works *Translatio s. Nicolai* and *Gesta principium polonorum*.[1]

Feature-based approaches are sometimes performant, but they are nearly always uninterpretable and therefore inflexible. Both sides of the problem—how exactly a given author produces texts of a particular type and how exactly do we ascertain that a given text was written by her—remain totally obscure even if the attribution is successful. More worryingly, there is often no clear way to estimate the degree of our uncertainty in the attribution and to check how well it is actually supported by the data.

Model-based approaches offer a way to overcome these limitations. These approaches regard text fragments by different authors as draws from a probability distribution characterizing this author's style. If we know the parameters of this distribution, we can directly estimate how likely it is that a given text was composed by this author.

In an ideal scenario, we should be able to estimate the probability that a given text was written by this author. It is easy to see, however, why this estimation is infeasible: in order for it to work, we need to construct a probability distribution over all possible author-text combinations, and we usually do not have access to the complete set of authors in any given language or genre.

Moreover, even if we had restricted the possible authors to some finite set, we still would not have had access to the joint probability of authors and textual fragments because they were not sampled in any meaningful sense. As a consequence of this issue, the current model-based approaches tend also to be partly feature based: the parameters of the model are estimated from the data, and their parameter vectors are then used as features in subsequent analysis.

The most important decision to be made here is what probability distribution to use for modelling feature distributions. The current instrument of choice is the multinomial distribution. This distribution models the probability of the event that after n trials each of which may result in K different outcomes the result will be equal to (p_1, p_2, …, p_k), where p_i is the number of trials ending in the outcome and all p_i's sum to n. Given a fixed vocabulary and a fixed text length, the probability distribution corresponding to a particular author computes the posterior probability or likelihood that she composed a given text sample based on counts of different words in it.

After choosing an appropriate model, the crucial task is to estimate its parameters, which will then be interpreted directly or serve as feature vectors. Scholars following the model-based approach [Gill et al. 2007; Gill, Swartz 2011] work in the Bayesian setting, and their methodology demands some assumptions about the probability distributions of values of these parameters (so-called *priors*). The multinomial distribution is parameterized by a vector of k values corresponding to probabilities of different outcomes in individual trials, which are assumed to be independent. These probabilities must sum to 1, and the logical prior to use here is the Dirichlet distribution. This distribution assigns probabilities to points in N-dimensional spaces with positive coordinates, whose values sum to 1.

[1] A radically different approach was employed in [Chaudhuri et al. 2018], who use syntactic features in order to distinguish between prose and verse. An interactive toolkit for syntax-centred stylometric analysis of Latin texts was recently published in [Bolt et al. 2019]. Other related work includes [Campbell et al. 2007] and [Stover, Kestemont 2016b].

The Dirichlet distribution is itself parameterized by a vector of k values usually denoted α. The Bayesian methodology demands that some values for these parameters are provided by the researcher. An 'uninformative' α with all elements equal to 1 or some other fixed value is usually chosen.

We focus on the use of Dirichlet distribution as a prior in detail because we think that it can be also profitably used as the primary modelling distribution. Ever since the work of Burrows, the research on stylometry made use of relative frequencies of function words. Burrows's approach was to normalize differences in relative frequencies of function words using z-scores.[2] It is possible, however, to treat the whole range of relative frequencies for a fixed vocabulary as a sample from the Dirichlet distribution characterizing the style of a particular author. This approach has the advantage that we directly model the quantity of interest and do not assume, as it happens when using the multinomial distribution, that words in the text were chosen independently of each other. In this study, we use maximum-likelihood inference to fit a Dirichlet distribution to the data.

The contributions of the paper are the following:

● We propose an interpretable, easily implementable, and very fast model-based probabilistic stylometric algorithm for authorship verification.

● We test the algorithm by assessing its ability to correctly attribute 21 different fragments for each author with a big enough œuvre in the reference corpus. We compare the performance of the algorithm with that of the random-forest classifier, an industry-standard ensemble method.

● We further 'stress test' the algorithm by limiting the number of training samples available to it from each author. The results of this procedure simultaneously assess the strength of the algorithm and show the degree of confusability between different authors in the corpus as well as general stylistic individuality. (E. g., fragments by Jerome and Marcus Terentius Varro remain fully identifiable even when the number of training samples is reduced to five while the success rate for Tacitus and Macrobius falls dramatically.)

The rest of the paper is organized as follows. Section 2 provides an overview of the corpus of Latin texts we used to test the new method. In § 3, we present the Dirichlet distribution in more detail, describe the algorithm used for estimating its parameters from the data, and the decision rule. The results of the analysis are presented in § 4. Section 5 presents our conclusions.

2. Data

We used all classical and medieval Latin prose writers whose body of work represented in the PHI5[3] and digilibLT[4] databases included more than 21 3000-word-long samples. This left us with 36 authors and 2395 text samples (i. e. 66.5 samples per author on average). See the full list of authors in § 4 below. We removed all non-alphabetic characters from the texts, lower-cased them, and replaced all *v*'s with *u*'s.

[2] A z-score for the relative frequency of a given word in a given text is computed by subtracting from it the mean relative frequency of this word in all texts from the corpus and dividing the result by the standard deviation of the same relative frequencies.

[3] https://latin.packhum.org.

[4] http://digiliblt.lett.unipmn.it/index.php?lang=en.

Following standard practices in stylometric research [Stover 2015; Stover, Kestemont 2016a], we selected 244 most-frequent genre and subject neutral words, predominantly function words, to construct vectors of relative frequencies (a 'non-round' number comes from the fact that we selected all common function words and then enlarged the list by adding topic independent most frequent content words; exploratory experiments with a smaller word set of 150 function words showed that it leads to a poorer performance).[5]

3. Methods

3.1. Dirichlet distribution

The Dirichlet distribution, also known as multivariate beta distribution, assigns probabilities to points in the open standard $(K-1)$-simplex, i. e., a set of points in a K-dimensional space whose coordinates are positive and sum to 1 (the simplex itself is thus a $(K-1)$-dimensional object because any set of $K-1$ coordinates uniquely determines the remaining one). The coordinates of a given point may be regarded as the probabilities of a K-way categorical event.

The probability density function of the Dirichlet distribution is defined as

$$f(x;\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}x_i^{\alpha_i-1}$$

where is a point in a standard $(K-1)$-dimensional simplex and α is the parameter vector.

For the purpose of this study, we take relative frequencies of selected function words in a text to be x vectors. In order to make the relative frequencies sum to 1, we collapse frequencies of all other words in a single additional category. The Dirichlet distribution demands that all elements of x be strictly positive. However, not all words are found in all text samples, which leads to zero frequencies. We normalize the data by adding very small positive constants to all relative frequencies in order to obviate this problem; this does not lead to issues with inference.

[5] et, in, non, ut, ad, cum, ab, sed, ex, si, de, etiam, enim, aut, ac, nec, per, atque, nam, uel, ne, quidem, autem, tamen, neque, uero, ita, iam, quoque, nihil, pro, modo, quia, quasi, inter, nisi, tunc, post, sic, igitur, tam, qua, ante, an, nunc, apud, magis, sine, ergo, at, deinde, ubi, dum, semper, minus, unde, contra, maxime, itaque, sicut, satis, denique, ob, simul, uti, sub, saepe, quamquam, numquam, ideo, propter, siue, quippe, prius, adhuc, quoniam, usque, inde, bene, sane, mox, item, super, quin, adeo, quamuis, cur, tamquam, postea, praeterea, potius, statim, uelut, postquam, supra, ceterum, certe, omnino, licet, forte, o, circa, rursus, tandem, diu, praeter, umquam, tot, ibi, hinc, haud, necesse, melius, paene, fere, namque, amplius, uix, scilicet, quum, iterum, aliquando, aduersus, seu, parum, plerumque, interim, prope, plus, intra, partim, olim, iuxta, ultra, male, quare, aliter, dolorem, fortasse, malis, primis, studio, agere, immo, quanto, domine, eiusdem, opera, oportet, publicam, tota, usus, aetatis, boni, locis, plurimum, potestate, saepius, antea, demum, dolore, imperatori, latine, malo, potuit, quadam, quondam, quosdam, sumus, dicuntur, diuina, lege, ordinem, postremo, regnum, solet, tribus, fama, patre, putat, hi, iubet, pluribus, quarum, sancti, solus, uera, uirtutes, uolunt, annis, dicam, dicta, domi, homo, ingenio, militum, studiis, uoce, a, aetate, castra, exercitum, genera, maior, summum, equidem, eundem, gratiam, loci, magnum, naturam, num, profecto, amicis, consules, etsi, honore, honorem, multos, quidquid, quisquam, dein, mali, mecum, sapientia, uiginti, accepit, cuiusque, exercitu, fuisset, plura, secum, domus, oratio, principis, uirtutis, iter, liberos, modi, ualde, alioquin, aqua, augusti, locus.

The Dirichlet distribution can be characterized by its mean

$$E[X_i] = \frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k}$$

and precision

$$s = \sum_k \alpha_k$$

When precision is high, samples from the distribution are likely to be near the expected value; when precision is small, they are distributed more diffusely. This allows for straightforward interpretation of fitted models. In our case, authors, whose associated Dirichlet distribution has high precision, are likely to have text samples that are similar to each other in their use of function words, while samples from authors with low precision are expected to be more varied. See [Bela et al. 2010] for more details.

3.2. The Dirichlet process and Latent Dirichlet Allocation

There is an important generalization of the Dirichlet distribution, the Dirichlet process. It allows researchers to model texts as mixtures of words taken from a number of different sources. The relationship between sources in this model is unequal: each next element is drawn proportionally to the number of elements from each source already in the set, thus giving rise to the 'rich get richer' model (a. k. a. the Chinese-restaurant process: elements from different sources are envisaged as patrons in a restaurant with an infinite number of tables preferring tables at which many people are already sitting).

The Dirichlet process is commonly used to model topic content of a collection of texts. As a result of this type of analysis, each text in the corpus is split into groups of words presumably connected to a different subject matter (such as *sports* or *politics*; the labels for topics are chosen by the researcher based on an analysis of word groups). Texts with similar proportions of elements from different classes are supposed to be thematically similar. Latent Dirichlet Allocation (LDA) [Blei et al. 2003] is a prominent example of this methodology.

LDA has been applied in a stylometric setting. One approach is to represent authors' styles as a distribution over topics extracted from a corpus and then compare these distributions for the purpose of analysis or attribution [Seroussi et al. 2014]. A more advanced approach is to jointly learn topical preferences and more fine-grained lexical preferences as 'the variations in choosing different words to convey a similar meaning introduce the lexical bias for an author to construct the document' [Ding et al. 2017: 4].

These models are very powerful and can be applied to large corpora of texts with very few assumptions. However, their reliance on topic modelling, with or without additional refinements, makes them vulnerable to the pitfalls of genre-induced similarities. In the present study, we have solid background knowledge about the texts' language and are confident in the discriminative power of our features (i.e., function words). Therefore, we resort to a simpler approach, one based on the basic Dirichlet distribution.

3.3. Inference and decision rule

There is no analytical solution for the problem of maximum-likelihood estimation of the parameters of a Dirichlet distribution from data. However, this distribution belongs to the exponential family of distributions, and therefore its log-likelihood function is convex. This makes it possible to use convex-optimization techniques for estimation. Several iterative algorithms for estimating α from data were proposed in [Minka 2012]. We adopted the approach formulated in his Equation 9:

$$\Psi(\alpha_k^{new}) = \Psi\left(\sum_k \alpha_k^{old}\right) + \frac{1}{N}\sum_i log x_{ik}$$

where N is the number of observations and Ψ is the digamma function.[6] We did not try to make a good first guess as to the seed values of α and set them all to $\frac{1}{K}$. This did not impact convergence.

36 authors with at least 21 3000-word-long samples were selected for the analysis. The analysis consisted of 21 iterations. Each iteration went as follows:

1. A test sample was randomly selected for each author; parameters of a Dirichlet distribution were estimated on the remaining 20+ samples. No test sample was selected twice (therefore, authors with the minimal number of samples, 21, had all their samples used as test samples).

2. For each test sample of each author, all author models were used to estimate the likelihood of this sample.

3. The author with the highest likelihood was selected as the best candidate author. The pair (real author, candidate author) was recorded for each sample.

Julia and R code together with the corpus are available at a public GitHub repository https://github.com/macleginn/dirichlet-stylometry-src.

4. Results

The results of identifying held-out test fragments (21 for each author, 756 altogether) using models fitted on all remaining training samples (only one held-out fragment was excluded each time) are shown in Table 1.[7] It should be noted

[6] Minka also proposes an algorithm for estimating the inverse of the digamma function, needed for the update, but we simply used the function `invdigamma` provided by the `SpecialFunctions.jl` module for the Julia programming language.

[7] The abbreviations stand for the following names: AGell: Aulus Gellius, AmmMarcell: Ammianus Marcellinus, Apul: Apuleius (spurious works excluded), Boeth: Anicius Manlius Severinus Boethius, CaelAurel: Caelius Aurelianus, Caes: Julius Caesar, Cels: Aulus Cornelius Celsus, Chalcid: Chalcidius, Cic: Cicero, Colum: Columella, Curt: Curtius Rufus, Digesta: the compiler of the *Digesta*, Diomed: Diomedes, FirmMatern: Firmicus Maternus (only the *Matheseos libri*), Hegesipp: Pseudo-Hegesippus, Hier: Jerome (only Biblical translations), LactPlac: Pseudo-Lactantius Placidus (the author of the extant scholia to Statius' *Thebaid*), Liv: Livy, Macrob: Macrobius, MarcellEmpir: Marcellus Empiricus, MartCap: Martianus Capella, NonMarcell: Nonius Marcellus, PlinMin: Pliny the Younger, PlinSec: Pliny the Elder, PompPorph: Pomponius Porphyrio,

that some of these texts are arguably not by a single author: the *Digesta* are a compilation from different juristic authorities of the 1st cent. BC through 4th cent. AD (mainly of the 2nd and 3rd cent. AD); Nonius Marcellus' *De compendiosa doctrina* and some parts of Macrobius' *Saturnalia* are basically long lists of quotations; Jerome's *Vulgate* is heavily indebted to the earlier *Vetus Latina*, which is probably not by a single translator (see [Kraus 2017: 121–124]), and some parts of the *Vulgate* may not be by Jerome at all (see [Rebenich 2002: 53]); Marcellus Empiricus' *De medicamentis* 'is a work of pure compilation which draws on the work of predecessors, above all Scribonius…, both Pliny [the Elder].... and the *Medicina Plinii*…, and Vindicianus…; Marcellus reproduces even the dedicatory epistles… and some statements in the first person of the authors he uses' [Langslow 2000: 66–67]; Servius (together with the additions of Servius Danielis, which are based on a different commentary), Pomponius Porphyrio, Pseudo-Lactantius Placidus, and the *Adnotationes super Lucanum* are sets of scholia, and such texts often reproduce verbatim numerous statements form earlier commentaries by different authors (see [Cameron 2004: 197–212, Zetzel 2018: 126–158]). Nevertheless, as shown by our experiments, the distribution of function words in these texts transcends their amalgamated nature and makes it possible to identify the compiler.

Values on the main diagonal indicate how many times the fragments were attributed correctly. The average attribution success rate for all authors is 93.9%, and for individual authors or text collections it ranges from 100% (AmmMarcell, Caes, Cic, Curt, and others) to 71% (Martianus Capella, the only author whose score is below 75%).

A high attribution success rate for a given author is not, however, to be interpreted straightforwardly as directly dependent on the originality of the author's style; for instance, there are hardly any reasons to doubt the originality of Apuleius' style, although his score is only 81%. The instances of confusion in Table 1 can virtually always be explained by affinities caused by imitation (e. g., Quintilian taken for Cicero; in total, at least 3 instances out of 46), chronological vicinity (e. g., Apuleius taken for Pliny the Younger; in total, 33 instances out of 46) or shared genre (e. g., Columella taken for Vegetius who also writes technical treatises; in total, 35 instances out of 46, 27 of them connected with technical style).[8] The last two kinds of affinities can actually also imply less readily recognizable imitation of one author by the other or imitation of a common source by both. Thus, a low attribution rate can mean three things: (i) the author is engaged in imitation of other authors in the corpus; (ii) the author is imitated by other authors in the corpus; (iii) the author is a typical exponent of a certain shared style represented in the corpus (generic, popular in a certain period, etc.). A high attribution rate, on the other hand, means that none of these conditions is satisfied and the author is less engaged in the network of connections established by the corpus, to an extent sufficient to let the algorithm reliably identify his style.

Quint: Quintilian, Sen: Seneca the Younger, SenMai: Seneca the Elder, Serv: Servius (as edited in [Thilo 1881–1887], i. e., Servius Danielis included), Suet: Suetonius, Symmach: Quintus Aurelius Symmachus, Tac: Tacitus, ValMax: Valerius Maximus, Varro: Marcus Terentius Varro, Veget: Vegetius, AdnSupLuc: the anonymous author (or compiler) of the *Adnotationes super Lucanum*.

[8] The only instance of confusion this classification does not account for is that between Suetonius and Macrobius.

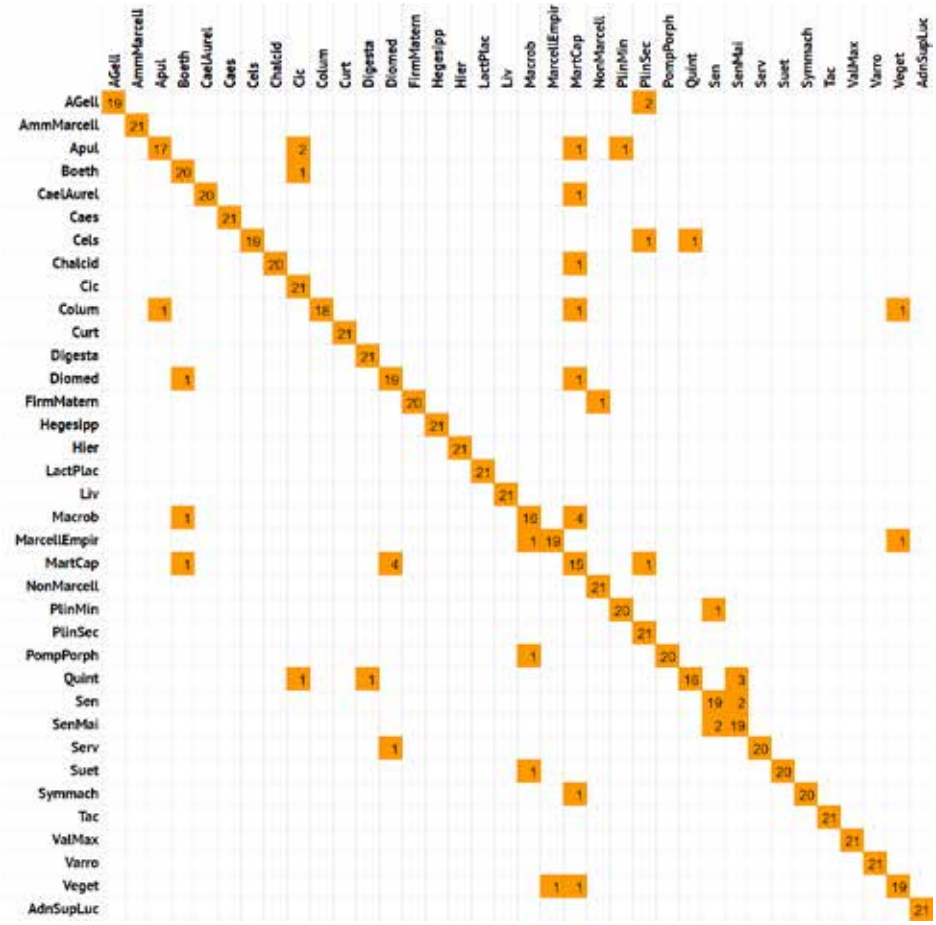

Table 1. *Confusion matrix for author identification of held-out samples by models trained on all available data. Empty cells indicate zeros*

Indeed, the authors our algorithm has particular problems with in this and the following experiments, like Martianus Capella or Macrobius, are all engaged in some forms of technical writing (a genre particularly well-represented in our corpus; for attempts at identifying features common to Latin technical texts that are connected with the use of functional words, see [Cousin 1943: 48–52; Langslow 2005: 297–298]). At the same time historians, who are also numerous in our corpus, almost always get very high attribution success rates, which is probably connected with the fact that their shared stylistic features seem to be rather connected with the general attitude towards innovation and archaism than with, say, particular formulaic expressions [Lebek 1970; Adams 2013: 260–267]. This kind of generic style does not prove a serious obstacle to our algorithm.

The author who gets the highest possible attribution success rates in all of our experiments is Caesar; among non-historians, the authors who succeed best are Jerome and Varro. Caesar and Varro can perhaps be seen as particularly non-typical

Table 2. *Precision of Dirichlet distributions fitted for each author*

Author	Precision
ValMax	1276
AmmMarcell	1203
AdnSupLuc	1178
Suet	1155
LactPlac	1120
Caes	1110
Curt	1098
Symmach	1047
AGell	1012
Hegesipp	1008
MarcellEmpir	1001
Serv	980
Liv	965
Tac	964
SenMai	962
Chalcid	961
Plin	953
NonMarcell	953
PompPorph	948
Sen	886
Cels	872
Veget	842
Colum	832
Digesta	831
Macrob	750
Varro	725
Cic	716
CaelAurel	703
Apul	691
Quint	665
PlinSec	620
Hier	591
MartCap	516
Diomed	513
FirmMatern	484
Boeth	325

representatives of their genres: Caesar is the earliest historian in the corpus and is not yet under the influence of Sallust, which, according to [Lebek 1970], was decisive for the later development of the style of Roman historians; Varro is also the earliest representative of the technical style in the corpus. Jerome's translations, despite their undoubted influence, are basically on their own in our corpus from the point of view of genre.

The precision of Dirichlet distributions (truncated to the integer part) fitted for different authors is shown in Table 2. Unsurprisingly, authors with high precision are easy to identify while authors characterized by low precision are among the most problematic for the classifier, especially when the number of training samples was restricted (see below). The upper part of the rating is again mostly occupied

by historians, while technical authors are concentrated at its bottom. Some authors with high attribution success rate get low precision though, notably Cicero, Varro, Jerome and the compiler of the Digesta; in these cases, the algorithm succeeds well in identifying texts by these authors even despite their stylistic non-uniformity.

A random-forest classifier [Breiman 2001] with 300 trees repeatedly fitted on all training data except for the same held-out fragments using the same features achieved a comparable, albeit slightly lower, attribution success of 91%. Importantly, it took 40 minutes to train the models and make predictions. Fitting 21 Dirichlet models for all authors took less than a minute, and the algorithm can be easily parallelized. This shows that the proposed approach is much more scalable.

The results of attributing the held-out fragments after restricting the training set for each author to 10 and 5 samples are shown in Tables 3 and 4 respectively. The accuracy with these training regimes dropped to 88% and 80.3% respectively showing the relative robustness of the methods to small training samples.

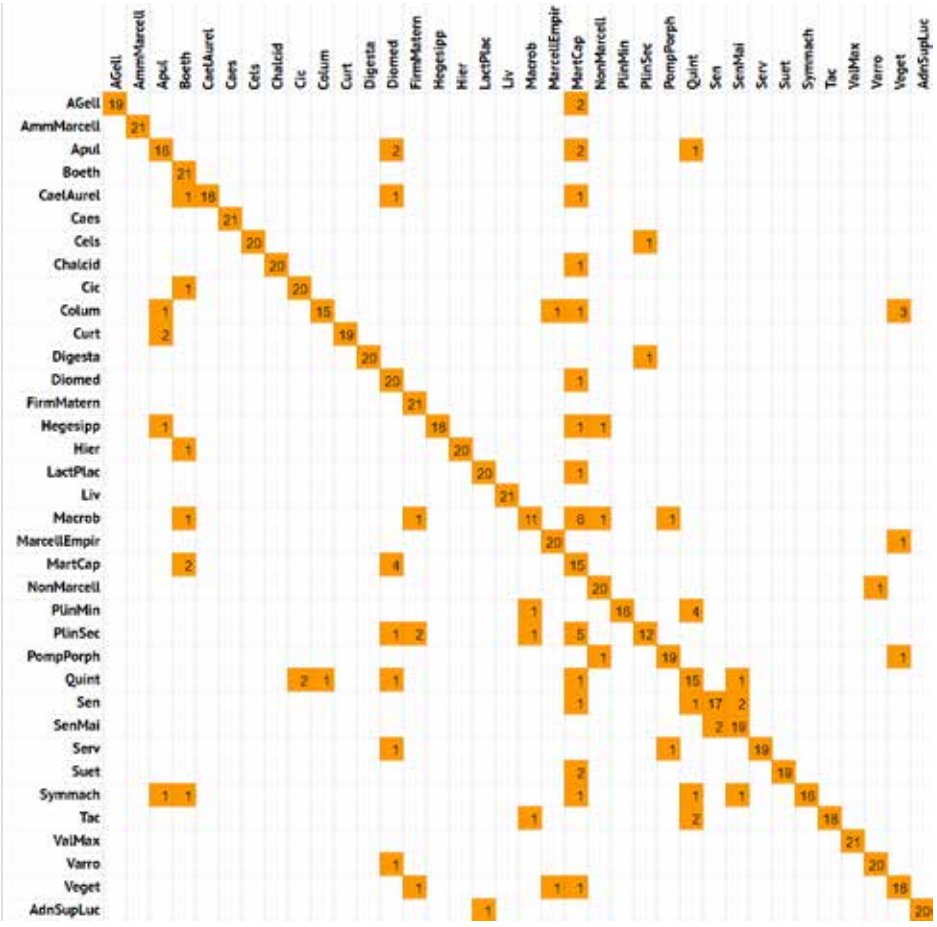

Table 3. *Confusion matrix for author identification of held-out samples by models trained on 10 samples for each author. Empty cells indicate zeros*

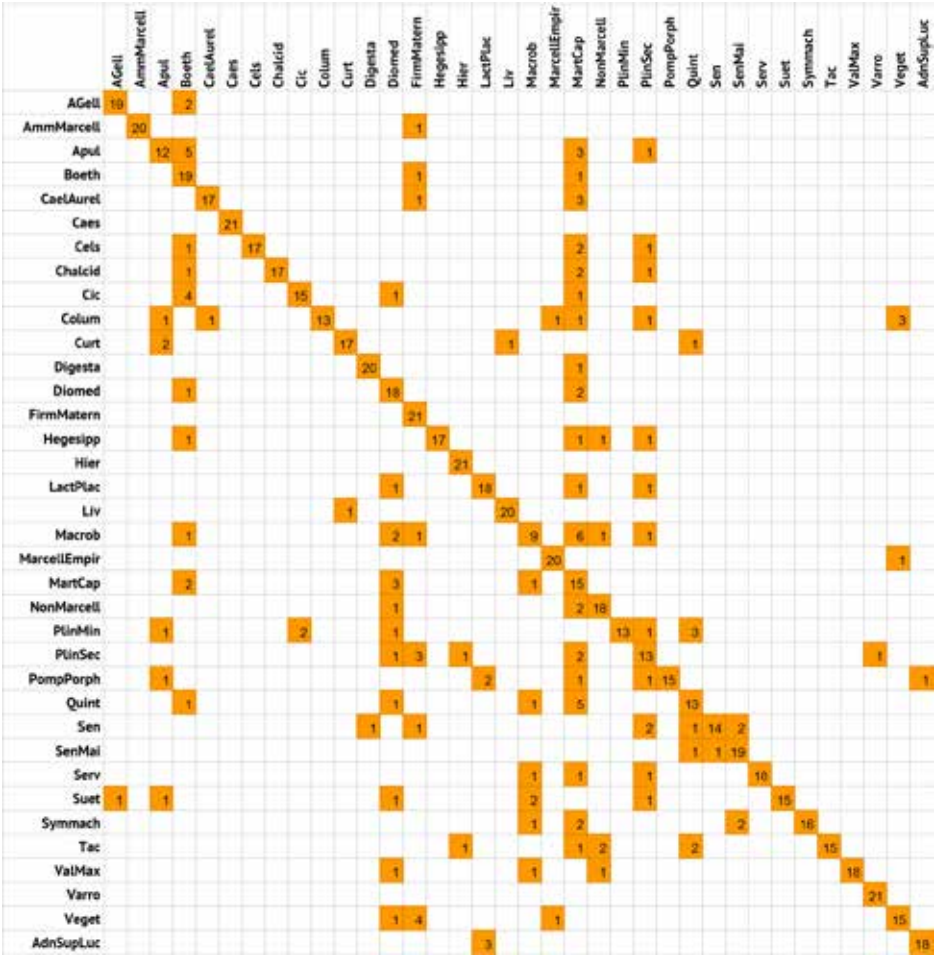

Table 4. *Confusion matrix for author identification of held-out samples by models trained on 5 samples for each author. Empty cells indicate zeros*

5. Conclusions

In this paper, we demonstrated the effectiveness of a simple and well-motivated method for author identification based on the Dirichlet distribution. It can be easily implemented in any general-purpose programming language (although using a special-purpose one, such as Julia or R, makes the implementation easier) and has good performance. Its practical application, e. g. in order to ascertain the authorship of a new fragment, consists of the following steps:

1. A table of relative frequencies of a set of words should be compiled based on a reference collection of text samples. Importantly, unlike when using Burrows's Delta, relative frequencies are not normalized, and there is no need to recompute them when the corpus is modified. This makes it possible to easily add new texts or subsample from the training data for testing purposes.

2. Models for different authors represented in the corpus are fit using the iterative method.

3. For a new sample, the likelihoods are computed estimating the confidence of different models that the sample was created by the respective author. The highest-likelihood model can be selected as the potential author or the results can be interpreted in another way.

The method does not provide an easy way to aggregate the results in order to ascertain, for instance, if the text was created by s o m e author represented in the sample or yet another one. However, this problem is by its nature not well defined statistically (unless there is some solid preliminary knowledge that can be encoded as a prior in a Bayesian setting), and we do not strive to solve it.

Furthermore, we do not claim that our approach provides the best possible results given the data. It may turn out that a well-tuned generalist model, such as a random-forest classifier or gradient-boosting machine, will outperform our algorithm on a particular dataset. However, the simplicity, performance, and interpretability of our approach make it a viable argument in favor of model-based approaches in stylometry.

References

Adams, J. N. (2013). *Social variation and the Latin language*. Cambridge Univ. Press.

Bela, A., Frigyik, A., & Gupta, M. (2010). *Introduction to the Dirichlet distribution and related processes*. Technical Report UWEETR-2010-0006. Department of Electrical Engineering, Univ. of Washington.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(4–5), 993–1022.

Bolt, T. J., Flynt, J. H., Chaudhuri, P., & Dexter, J. P. (2019). A stylometry toolkit for Latin literature. In S. Padó, R. Huang (Eds.). *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations* (pp. 205–210). Association for Computational Linguistics.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267–287.

Cameron, A. (2004). *Greek mythography in the Roman world*. Oxford Univ. Press.

Campbell, G., Corns, T. N., Hale, J. K., & Tweedie, F. J. (2007). *Milton and the manuscript of* De Doctrina Christiana. Oxford Univ. Press.

Chaudhuri, P., Dasgupta, T., Dexter, J. P., & Iyer, K. (2018). A small set of stylometric features differentiates Latin prose and verse. *Digital Scholarship in the Humanities, 34*(4), 716–729.

Cousin, J. (1943). Les langues spéciales. In *Mémorial des études latins: Publié à l'occasion du vingtième anniversaire de la Société de la Revue des études latines, offert par la Société à son fondateur J. Marouzeau* (pp. 37–54). Les belles lettres. (In French).

Ding, S. H. H., Fung, B. C. M., Iqbal, F., & Cheung, W. K. (2017). Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics, 49*(1), 107–21.

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. *The R Journal, 8*(1), 107–121.

Gill, P. S., & Swartz, T. B. (2011). Stylometric analyses using Dirichlet process mixture models. *Journal of Statistical Planning and Inference, 141*(11), 3665–3674.

Gill, P. S., Swartz, T. B., & Treschow, M. (2007). A stylometric analysis of King Alfred's literary works. *Journal of Applied Statistics, 34*(10), 1251–1258.

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

Holmes, D. I., & Kardos, J. (2003). Who was the author? An introduction to stylometry. *Chance, 16*(2), 5–8.

Kabala, J. (2020). Computational authorship attribution in medieval Latin corpora: The case of the Monk of Lido (ca. 1101–08) and Gallus Anonymous (ca. 1113–17). *Language Resources and Evaluation, 54*, 1–32.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., & Daelemans, W. (2016). Authenticating the writings of Julius Caesar. *Expert Systems with Applications, 63*, 86–96.

Koppel, M., & Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology, 65*(1), 178–187.

Kraus, M. A. (2017). *Jewish, Christian, and classical exegetical traditions in Jerome's translation of the Book of Exodus: Translation technique and the vulgate*. Brill.

Langslow, D. R. (2000). *Medical Latin in the Roman Empire*. Oxford Univ. Press.

Langslow, D. R. (2005). 'Langues réduites au lexique?' The languages of Latin technical prose. In T. Reinhardt, M. Lapidge, J. N. Adams (Eds.). *Aspects of the language of Latin prose* (pp. 287–302). Oxford Univ. Press.

Lebek, W. D. (1970). *Verba prisca: Die Anfänge des Archaisierens in der lateinischen Beredsamkeit und Geschichtsschreibung*. Vandenhoeck und Ruprecht. (In German).

Minka, T. P. (2012). *Estimating a Dirichlet distribution*. https://tminka.github.io/papers/dirichlet.

Mosteller, F., & Wallace, D. L. (1964). *Applied Bayesian and classical inference: The case of the Federalist Papers*. Addison-Wesley.

Rebenich, S. (2002). *Jerome*. Routledge.

Seroussi, Y., Zukerman, I., & Bohnert, F. (2014). Authorship attribution with topic models. *Computational Linguistics, 40*(2), 269–310.

Stover, J. A. (Ed. & Trans.) (2015). *A new work by Apuleius: The lost third book of the* De Platone. Oxford Univ. Press.

Stover, J. A., Winter, Y., Koppel, M., & Kestemont, M. (2016). Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the Association for Information Science and Technology, 67*(1), 239–242.

Stover, J. A., & Kestemont, M. (2016a). Reassessing the Apuleian corpus: A computational approach to authenticity. *The Classical Quarterly, 66*(2), 645–72.

Stover, J. A., & Kestemont, M. (2016b). The authorship of the *Historia Augusta*: Two new computational studies. *Bulletin of the Institute of Classical Studies, 59*(2), 140–157.

Thilo, G. (Ed.). (1881–18877). *Servii grammatici qui feruntur in Vergilii carmina commentarii*. Teubner. (In Latin).

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks, 10*(5), 988–999.

Zetzel, J. E. G. (2018). *Critics, compilers, and commentators: An introduction to Roman philology, 200 BCE — 800 CE*. Oxford Univ. Press.

* * *

Информация об авторах

Дмитрий Сергеевич Николаев
кандидат филологических наук
старший научный сотрудник,
Лаборатория теоретической фольклористики, Школа актуальных гуманитарных исследований,
Российская академия народного хозяйства и государственной службы при Президенте РФ
Россия, 119606, Москва, пр-т Вернадского, д. 82
Тел.: +7 (499) 956-96-47
научный сотрудник,
факультет лингвистики,
Стокгольмский университет
Швеция, SE-106 91, Стокгольм
Тел.: +46 (8) 16-20-00
✉ *dnikolaev@fastmail.com*

Михаил Владимирович Шумилин
кандидат филологических наук
старший научный сотрудник,
Лаборатория античной культуры,
Школа актуальных гуманитарных исследований, Российская академия народного хозяйства и государственной службы при Президенте РФ
Россия, 119571, Москва, пр-т Вернадского, д. 82
Тел.: +7 (499) 956-96-47
старший научный сотрудник,
отдел античной литературы,
Институт мировой литературы им. А. М. Горького РАН
Россия, 121069, Москва, ул. Поварская, д. 25а
Тел.: +7 (495) 690-50-30
✉ *mvlshumilin@gmail.com*

Information about the authors

Dmitry S. Nikolaev
Cand. Sci. (Philology)
Senior Researcher,
Center for Theoretical Folklore Studies, School for Advanced Studies in the Humanities,
The Russian Presidential Academy of National Economy and Public Administration
Russia, 119606, Moscow, Prospekt Vernadskogo, 82
Tel.: +7 (499) 956-96-47
Researcher,
Department of Linguistics,
Stockholm University,
Sweden, SE-106 91 Stockholm
Tel.: +46 (8) 16-20-00
✉ *dnikolaev@fastmail.com*

Mikhail V. Shumilin
Cand. Sci. (Philology)
Senior Researcher,
Laboratory of Classical Studies, School for Advanced Studies in the Humanities, The Russian Presidential Academy of National Economy and Public Administration
Russia, 119571, Moscow, Prospekt Vernadsogo, 82
Tel.: +7 (499) 956-96-47
Senior Researcher,
Department of Classical Literature,
A. M. Gorky Institute of World Literature of the Russian Academy of Sciences
Russia, 121069, Moscow, Povarskaya Str., 25a
Tel.: +7 (495) 690-50-30
✉ *mvlshumilin@gmail.com*