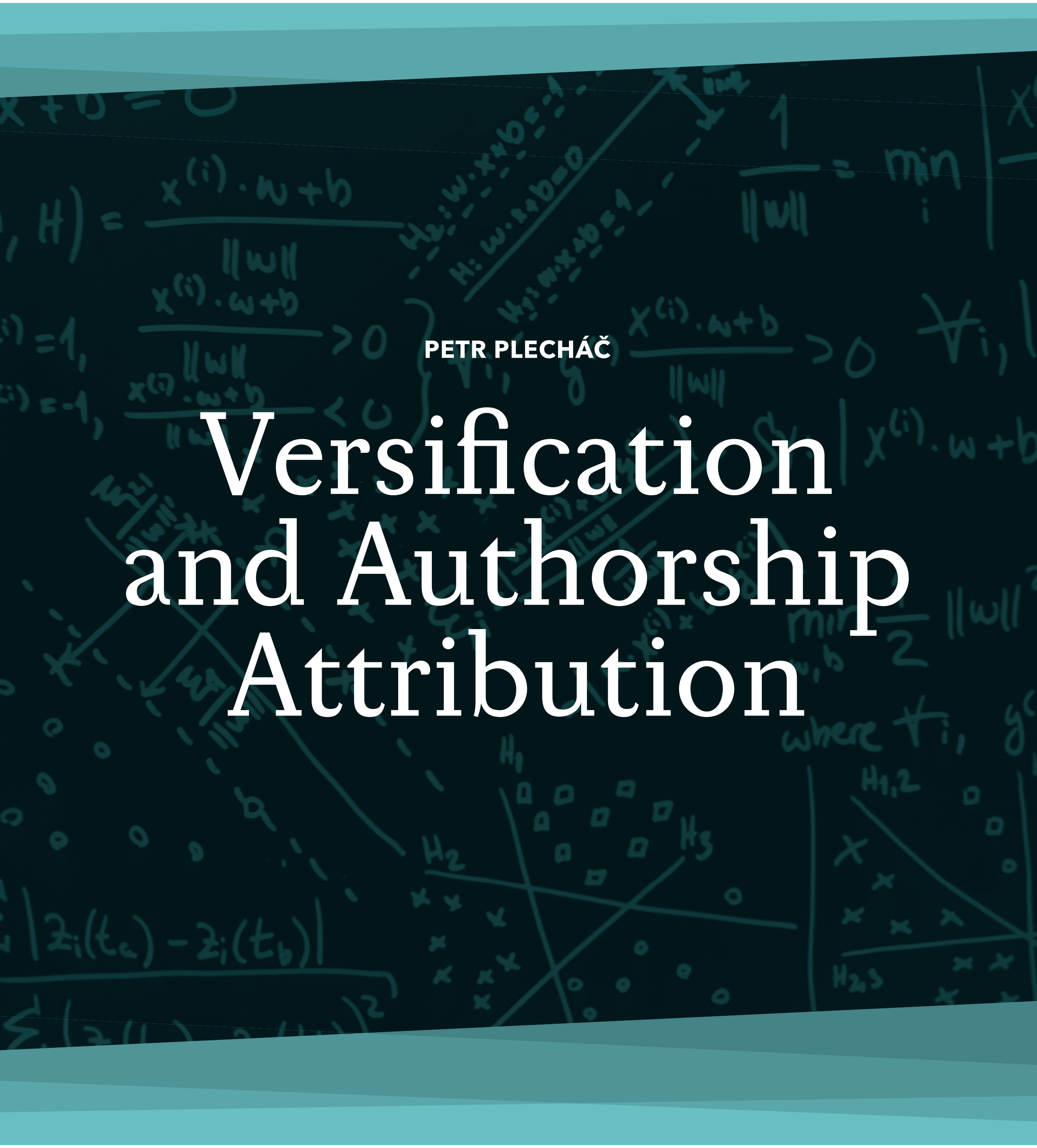
CZECH LITERATURE STUDIES



PETR PLECHÁČ

Versification and Authorship Attribution

INSTITUTE OF CZECH LITERATURE
KAROLINUM PRESS

Versification and Authorship Attribution

Petr Plecháč

Original manuscript reviewed by Mike Kestemont (University of Antwerp)
and Igor Pilshchikov (University of California, Los Angeles).

INSTITUTE OF CZECH LITERATURE is a part of the Czech Academy of Sciences
Na Florenci 1420/3, 110 00 Prague 1, Czech Republic
www.ucl.cas.cz

KAROLINUM PRESS is a publishing department of Charles University
Ovocný trh 560/5, 116 36 Prague 1, Czech Republic
www.karolinum.cz

Language review by Debra Shulkes; technical language review by Benjamin Nagy
Cover and graphic design DesignIQ
Set in the Czech Republic by Karolinum Press
First edition

Cataloguing-in-Publication Data is available from the National Library of the Czech Republic.

This study is the result of research funded by the Czech Science Foundation as part of project GA ČR 17-01723S.

This publication was created with the support of Research Development Program RVO 68378068 and published with support from the Czech Academy of Sciences.

https://doi.org/10.14712/9788024648903

ISBN 978-80-7658-027-5 (Institute of Czech Literature of the Czech Academy of Sciences)
ISBN 978-80-246-4871-2 (Karolinum Press)
ISBN 978-80-7658-028-2 (pdf, Institute of Czech Literature of the Czech Academy of Sciences)
ISBN 978-80-246-4890-3 (pdf, Karolinum Press)

Contents

Introduction 7

Previous Publications 8

Data and Code 8

HTML version 8

1 Quantitative Approaches to Authorship Attribution 9

1.1 Origins of Stylometry 9

1.2 Searching for the "Golden Feature" 12

1.3 Multivariate Analyses 13

1.4 Support-Vector Machines 19

1.5 Versification-Based Attribution 32

1.6 Summary 34

2 Versification Features 36

2.1 Rhythm 36

2.2 Rhyme 41

2.3 Euphony 42

3 Experiments 43

3.1 Data 43

3.2 Versification-Based Attribution 47

3.3 Comparison with Lexicon-Based Models 56

3.4 Summary 63

4 Applications 69

4.1 *The Two Noble Kinsmen* 69

4.2 The Case of (Pseudo-)Batenkov: Towards a Formal Proof of Literary Forgery (co-authored by Artjoms Šeļa) 80

References 92

Introduction

Contemporary stylometry is one of the fastest-growing fields in the computational study of literature. In recent years, a number of textual characteristics and machine learning techniques have proven highly accurate in distinguishing the texts of different authors. Many of these features like word and character *n*-gram frequencies amount, however, to what is known as statistical "rare events", or more precisely, a large number of rare events (LNRE). As a result, their analysis calls for fairly large text samples consisting of thousands or tens of thousands of words. Most theoretical studies in stylometry therefore focus on long novels. Poetry is usually omitted although we might expect to find many more cases of disputed authorship among poetic works.

At the same time, poetry has a number of specific versification features that are essentially Boolean or open to only a limited number of values. Some of these—stanza length and rhyme scheme, for example—are subject to the author's conscious selection and so unsuitable for authorship recognition. In contrast, others like the preference for certain rhythmic configurations or sound frequencies in rhyme may be outside the author's rational control. Although these characteristics have traditionally been recognised as author-specific (or at least period-specific), they have rarely featured in authorship attribution studies.

The goal of this book is to examine the applicability of these versification features to authorship attribution projects. To this end, I draw on poetic corpora in three different languages (Czech, German and Spanish) and apply this approach to two real-world cases of disputed authorship.

Chapter 1 gives a brief history of quantitative methods of authorship attribution with special attention to the methods used in this book.

Chapter 2 highlights different ways to capture versification features.

Chapter 3 describes experiments with versification-based attribution and compares the methods commonly used in stylometry.

Finally, Chapter 4 applies these findings to two actual cases of ambiguous authorship involving English- and Russian-language texts respectively. In the first case,

https://doi.org/10.14712/9788024648903.1

I attempt to determine which parts of the verse play *The Two Noble Kinsmen* were written by William Shakespeare and which were the work of his co-author, John Fletcher. In the second, working together with Artjoms Šeļa, I investigate the potential forgery of numerous 19th-century Russian poems that were originally attributed to Gavriil Stepanovich Batenkov. These poems first appeared in the 1978 edition of the poet's collected works, which was compiled by an established literary scholar—the main suspect in this intrigue.

Previous Publications

Chapter 1 expands on the opening sections of *Versification and authorship attribution. Pilot study on Czech, German, Spanish, and English poetry* (Plecháč, Bobenhausen and Hammerich 2018).

Czech versions of Chapters 1, 2 and 3 were submitted as part of my PhD thesis at Charles University in Prague, Czech Republic in 2019.

Data and Code

The data and code required to reproduce the analyses in this book can be found at <https://doi.org/10.5281/zenodo.4555250>.

HTML version

From early 2022, this book will also be available online at <https://versologie.cz/versification-authorship>.

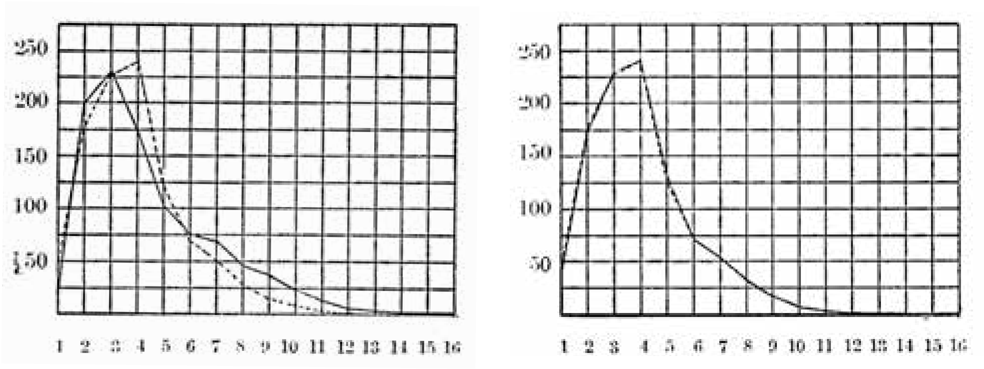
1 Quantitative Approaches to Authorship Attribution

1.1 Origins of Stylometry

Many scholars (e.g. Holmes 1998; Juola 2006) trace the origins of stylometry to several passages in a letter written by the British mathematician Augustus De Morgan to Reverend W. Heald on August 18, 1851 (De Morgan 1851/1882). After considering how to distinguish the Pauline epistles actually written by St. Paul from those written by other author(s), De Morgan mused that the average word length measured by the number of characters might give some clue: "If St. Paul's epistles which begin with Παυλος gave 5.428 and the Hebrews gave 5.516, for instance, I should feel quite sure that the *Greek* of the Hebrews (passing no verdict on whether Paul wrote in Hebrew and another translated) was not from the pen of Paul" (De Morgan 1851/1882: 216; emphasis in the original). Later he complained: "If scholars knew the law of averages as well as mathematicians, it would be easy to raise a few hundred pounds to try this experiment on a grand scale" (De Morgan 1851/1882: 216).

In fact, it was not until the end of the 19th century that the American physicist Thomas Corwin Mendenhall raised the money for this experiment. In an initial article entitled "The Characteristic Curve of Composition" (1887), Mendenhall suggested ignoring averages and dealing with overall word length distribution instead. Eventually, thanks to the support of a benefactor, August Hemenway, he applied this method to a real-world case of disputed authorship. The results of that experiment were published in the article "A Mechanical Solution to a Literary Problem" (1901). There, Mendenhall compared the shape of a curve determined by the relative frequencies of words of different lengths in works ascribed to William Shakespeare with equivalent curves for works by Francis Bacon and Christopher Marlowe (FIG. 1.1). Based on the similarities and differences, he cautiously concluded that while Bacon had not written the works in question, there was strong evidence that Marlowe had (Mendenhall 1901: 104–105). The discrepancies between the curves for Shakespeare and Bacon were, however, later found to be due to the comparison of verse texts by the former with non-verse texts by the latter (see Williams 1975).

https://doi.org/10.14712/9788024648903.2



(a) Texts ascribed to Shakespeare (dashed line) and texts by Bacon (solid line).

(b) Texts ascribed to Shakespeare (dashed line) and texts by Marlowe (solid line almost covering dashed line).

FIG. 1.1: Relative frequencies (per thousand) of word lengths measured by number of characters; source: Mendenhall 1901: 104 (facsimile).

Independently of Mendenhall, the American mathematician William Benjamin Smith had also been employing quantitative methods in the 1880s. In his article "Curves of Pauline and Pseudo-Pauline Style", published under the pen name Conrad Mascol (1888a; 1888b), he, like De Morgan, considered the authorship of the Pauline epistles. In line with Mendenhall, he took the shape of the curves representing various textual features (e.g. the average number of words or prepositions per page) to be a criterion. On comparing the curves for epistles generally agreed to be written by St. Paul with those of doubtful authorship, Smith concluded that the author of the former had probably not written the latter. Significantly, he also stressed that the key consideration when selecting features should be their topic independence.[1] This principle, though now taken for granted, was not generally accepted until the mid-20th century, as we will see in Section 1.2.

A third pioneering work usually mentioned in this field is an article by Lucius Adelno Sherman (1888) that was probably also conceived independently of Mendenhall's studies.[2] It analysed the average sentence length measured by the number of

[1] Smith wrote: "When we now ask, What are the elements of style to be considered? The answer must be: All such as are affected not at all, or apparently and comparatively very little, by the subject-matters of discourse" (Mascol 1888a: 456).

[2] Grzybek (2014) notes, however, that Sherman may have been inspired by a response to Mendenhall's initial article that was published in an 1887 issue of *Science*. Its author observed: "There are other characteristics of writers equally susceptible of treatment by the statistical and graphical method, in

words in the work of novelists writing in English. Still Sherman did not highlight the possibility of using this metric for authorship recognition.

Outside of these studies, there is, however, another branch of stylometry which, although only sporadically recognised by scholars (Grzybek 2014 and Grieve 2005 rank among the exceptions), dates back some 100 years before Mendenhall's first article and more than 60 years before De Morgan's letter. This concerns the attributions of Shakespearean scholars based on the quantification of rhythm and rhyme.

One of the earliest examples of this approach can be found in a study by Edmond Malone (1787/1803) which proposed that none of the three parts of the play *Henry VI* had actually been written by Shakespeare. Malone's arguments were based, among other things, on attention to versification: he argued that there were far fewer rhymes and enjambments in the texts in question than in other works by Shakespeare.

Another instance can be seen in a comment by the scholar Henry Weber about the play *The Two Noble Kinsmen* (1812), which was first published in 1634 as a collaborative work by William Shakespeare and John Fletcher (see Section 4.1 for details). Weber worked out a scene-by-scene division of authorship between Shakespeare and Fletcher based on the frequencies of certain line endings among other factors:

> Taking an equal number of lines in the different parts which are attributed to Shakespeare and to Fletcher, the number of female, or double terminations in the former, is less than one to four; on the contrary, in the scenes attributed to Fletcher the number of double or triple terminations is nearly three times that of single ones. (Weber 1812: 166)

Decades later, James Spedding (1850) used the same metric to arrive at a theory of joint authorship by Shakespeare and Fletcher that he also applied to *Henry VIII*.

The real rise of versification-oriented stylometry did not come, however, until the 1870s and 1880s after the founding of the *New Shakspere Society*.[3] In the first volume of their *Transactions,* one Society member, John Kells Ingram (1874) suggested dividing unstressed blank verse endings into "light endings" and "weak endings"[4] and using

which their personal peculiarities differ more widely, and which are therefore more characteristic than the habitual selection and use of long or short words. For example: it seems to me that the length of the sentence is such a peculiarity" (Eddy 1887: 297).

[3] Concerning its name, the Society's members maintained: "This spelling of our great Poet's name is taken from the only unquestionably genuine signatures of his that we possess, the three on his will, and the two on his Blackfriars conveyance and mortgage." (Furnivall 1874a: 6).

[4] Ingram described these two forms as follows: "It is evident that amongst what have been called as a class weak endings, there are different degrees of weakness. […] There are *two* such degrees, which require to be discriminated, because on the words, which belong to one of these groups the voice can

the ratio of instances of the two to support Spedding's attribution of *Henry VIII*. Ingram himself called this method the "weak-ending test". Other members proposed (or adopted) and applied several other such verse tests designed to distinguish Shakespeare's works from those of other authors based on the prevalence of particular features. These included the "rhyme test" (for rhymed lines), the "stopt-line test" (for enjambment), the "middle-syllable test" (for extra-metrical syllables at the end of the first half-line) and the "caesura test" (for word breaks after the sixth syllable in alexandrines).[5]

Many of these attributions by New Shakspere Society members were later proven wrong owing to the simplistic nature of their methods or errors in their source data (Grieve 2005: 6). Even so, they are an important part of the history of stylometry and should not be neglected.

1.2 Searching for the "Golden Feature"

The works of George Kingsley Zipf seem to have inspired a new era in the development of 20th-century stylometry (see Koppel, Schler and Argamon 2009: 4–5). The formulation of Zipf's law (1932), which states that all natural language texts follow the same rank-frequency word distribution, likely encouraged scholars to rethink the possibilities for authorship attribution. This meant finding a similar textual feature that would remain stable across the works of one author while differing in those of other authors.

Of great influence in this period were the stylometric works of George Udny Yule, who initially proposed using sentence length measured by the number of words (Yule 1939). Unlike Sherman (see Section 1.1), Yule considered not only average values but also other distribution characteristics. These included the median, the $Q_{0,25}$ and $Q_{0,75}$ quartiles, the interquartile range and also—since sentence length generally tends to follow a positively skewed log-normal distribution—the decile $Q_{0,9}$.

Just a couple of years later, Yule's book *The Statistical Study of Literary Vocabulary* (1944) introduced a new metric designed to capture vocabulary richness. He defined that measure as follows:

to a certain small extent dwell, whilst the others are so essentially *proclitic* in their character […] that we are forced to run them, in pronunciation no less than in a sense, into the closest connection with the opening words of the succeeding line. The former may with convenience be called 'light endings', whilst to the latter may be appropriated the name (hitherto vaguely given to both groups jointly) of 'weak endings'" (Ingram 1874: 447; emphasis in original).

[5] See Fleay 1874a, 1874b, 1874c, 1874d; Furnivall 1874b, 1874c.