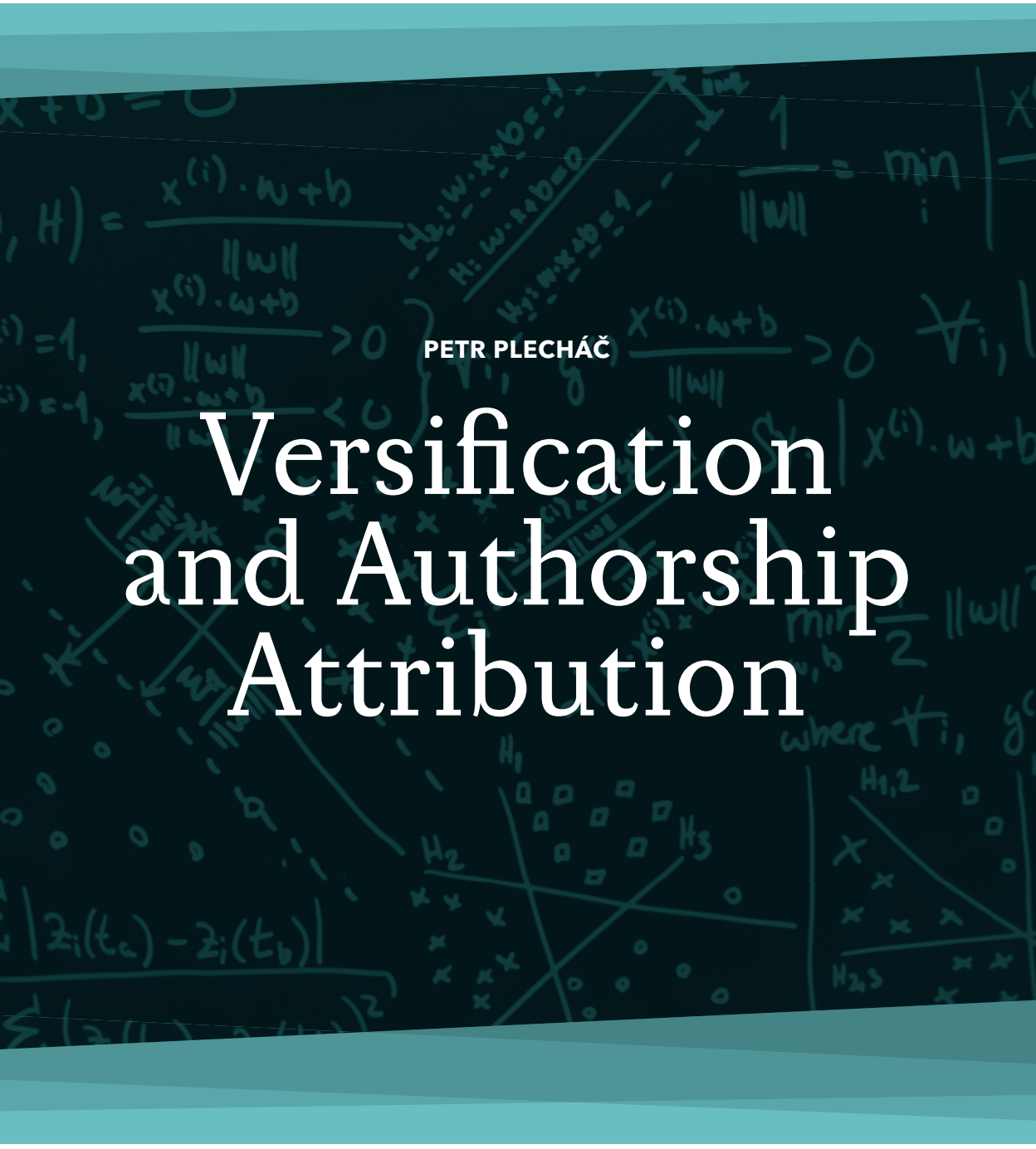
CZECH LITERATURE STUDIES



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

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›.

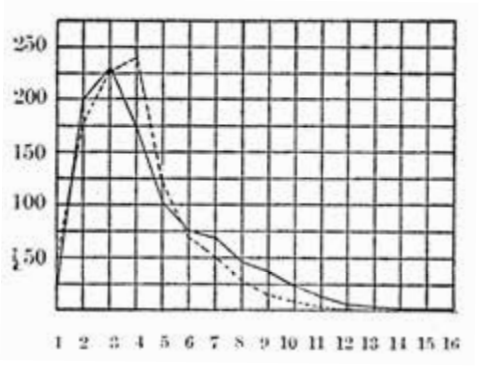
1 Quantitative Approaches to Authorship Attribution

1.1 Origins of Stylometry

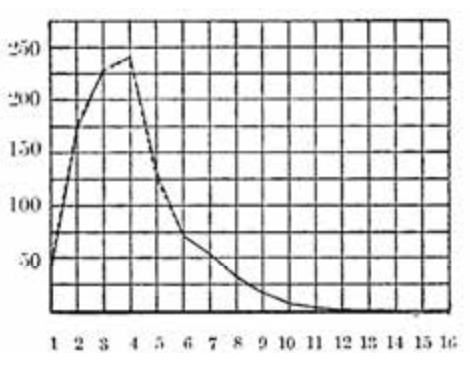
Many scholars (e.g. Holmes 1998; Juola 2006) trace the origins of stylometry to several passages in a letter written by the British mathematician Augustus De Morgan to Reverend W. Heald on August 18, 1851 (De Morgan 1851/1882). After considering how to distinguish the Pauline epistles actually written 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