Akademie věd
České republiky

Teze disertace
k získání vědeckého titulu „doktor věd“
ve skupině vědy filologické

QUANTITATIVE ANALYSES OF POETIC TEXTS

Komise pro obhajoby doktorských disertací v oboru bohemistika

Mgr. Petr Plecháč, Ph.D. & Ph.D.

Ústav pro českou literaturu AV ČR, v. v. i.

Praha 2022

Contents

Introduction

This thesis comprises articles on corpus verse studies and stylometry published during the last decade.

The first group of articles deals with problems of building an automatically annotated corpus of poetic texts, that is, the machine recognition of meter and rhyme. Back in 2011, when Robert Kolár (Ibrahim) and I began considering the options for building an automatically annotated database of Czech poetry, this was seen as a challenging task. At the time, no such work had been undertaken and the few theoretical studies available on the matter proposed simple rule-based algorithms for German and Russian versification (Bobenhausen and Gehl 2009; Pilschchikov et al. 2011). In line with these studies, our first attempts (described in *Towards the Automatic Analysis of Czech Verse*, 2011) had a purely philological basis and followed the rule-based system proposed by Miroslav Červenka (2006). Despite that system's simplicity and its reliance on introspection rather than data analysis, it turned out to be robust enough for our purposes. We, thus, used it to build the pilot version of the Corpus of Czech Verse (CCV)– one of the pioneering efforts to establish an automatically annotated database of poetic texts consisting of more than 80,000 poems from the late 18th century to the beginning of the 20th century.

Since then, many teams of researchers around the globe have followed this same path and corpus-based verse scholarship has become an established field of study. The following are just a few of these projects: simultaneously with the creation of CCV, a team in Freiburg succeeded in building an extensive corpus of German poetry (https://metricalizer.de/); a corpus of Spanish sonnets of the “golden age” was compiled at the University of Alicante (https://github.com/bncolorado/CorpusSonetosSigloDeOro); and a corpus of French poetry was developed at the University of Caen (https://crisco2.unicaen.fr/verlaine/).

The second article in this group, *Czech Verse Processing System KVĚTA—Phonetic and Metrical Components*, 2016) proposes a more advanced method of meter recognition in Czech poetry based on supervised machine learning. And the third article, (*A Collocation-Driven Method of Discovering Rhymes (in Czech, English, and French Poetry)*, 2018) describes a language-independent unsupervised method of rhyme discovery. Both these studies illustrate a general methodological shift in recent years from rule-based approaches to meter and rhyme recognition (Bobenhausen 2011; Bobenhausen and Hammerich 2015; Navarro-Colorado 2015; Navarro-Colorado et al. 2016; Pilschchikov et al. 2011) to more robust and accurate machine learning models (Haider 2021; Zabaleta 2017) .

The second group of articles explores the options for using these automatically detected versification features for automatic authorship recognition. The last two decades have seen rapid advances in work on the machine-driven authorship recognition of literary texts: from the influential studies of John F. Burrows (2002) , which proposed the famous Delta measure, to more advanced machine-learning based approaches (see Savoy (2020)). Many textual features (e.g. word, character *n*-gram and POS-tag frequencies) have been proposed for this purpose. Versification has, however, (almost) never been included among them. The articles in this group, thus, propose and apply a novel approach to authorship recognition of poetic texts–the versification-based method.

The first article tests the performance of versification-based authorship recognition with verse corpora in four different languages (*Versification and Authorship Attribution. A Pilot Study on Czech, German, Spanish, and English Poetry*, 2018). The second article (*Assessing the Reliability of Stress as a Feature of Authorship Attribution in Syllabic and Accentual Syllabic Verse*, 2019) examines potential relations between recognition accuracy and versification typology. The final two articles apply versification-based authorship recognition to two cases involving disputed authorship: The first concerns the verse play *Henry VIII*, which was originally published under the name of William Shakespeare, but is believed by many to contain parts that were actually written by John Fletcher or perhaps even by other authors (*Relative Contributions of Shakespeare and Fletcher in Henry VIII: An Analysis Based on Most Frequent Words and Rhythmic Patterns*, 2021). The second relates to the authorship of the Old English poem *Beowulf* (*Beowulf Single-Authorship Claim is Unsupported*, [accepted]).

Generally speaking, the articles are reprinted here in the form in which they were originally published. Additional editing has been limited to the unification of citation and graphic styles and correction of several typos. However, in keeping with Muphry's law, these adjustments may have led to the introduction of new typos, and I apologize in advance for those errors.

- *Towards the Automatic Analysis of Czech Verse* was published in the book **Formal Methods in Poetics, 2011 (RAM Verlag)**
- *Czech Verse Processing System KVĚTA—Phonetic and Metrical Components* was published in ***Glottotheory 7.2, 2016 (De Gruyter)***
- *A Collocation-Driven Method of Discovering Rhymes (in Czech, English, and French Poetry* was published in the book ***Taming the Corpus. From Inflection and Lexis to Interpretation, 2011 (Springer)***
- *Versification and Authorship Attribution. A Pilot Study on Czech, German, Spanish, and English Versification* was published in ***Studia Metrica et Poetica 5.2, 2018 (University of Tartu)***. This article was co-authored by Klemens Bobenhausen and Benjamin Hammerich (both Metricalizer, Freiburg im Brsg., Germany). Parts of it were also submitted as part of my second PhD thesis (Authorship Attribution of Poetic Texts, Charles University, 2019)
- *Assessing the Reliability of Stress as a Feature of Authorship Attribution in Syllabic and Accentual Syllabic Verse* was published in the Proceedings of the ***Quantitative Approaches to Versification, 2020 (ICL, Czech Academy of Sciences)***. This study was co-authored by David J. Birnbaum (University of Pittsburgh, Pennsylvania)
- *Relative Contributions of Shakespeare and Fletcher in Henry VIII: An Analysis Based on Most Frequent Words and Most*

Frequent Rhythmic Patterns was published in ***Digital Scholarship in the Humanities 36.2, 2021 (Oxford University Press)***. The article builds on a chapter on the play's authorship that was part of my second PhD thesis (*Authorship Attribution of Poetic Texts*, Charles University, 2019).

- *Beowulf Single-Authorship Claim is Unsupported* was published in ***Nature Human Behaviour 5, 2021 (Springer Nature)*** and co-authored by Andrew Cooper (University of Stockholm), Benjamin Nagy (University of Adelaide, South Australia), and Artjoms Šeļa (University of Tartu, Estonia)

Extended Abstracts

Towards the Automatic Analysis of Czech Verse

Robert Ibrahim and Petr Plecháč

ICL, Czech Academy of Sciences, Prague, Czech Republic

Formal Methods in Poetics, 2011 (RAM Verlag)

This paper presents a project which aims to develop software suitable for the metrical (-rhythmical) analysis of Czech verse.

We represent each line in a poem as a sequence of the following syllable-types: (a) initial syllable of a polysyllabic unit, (b) non-initial syllable of a polysyllabic unit, (c) monosyllabic preposition, or (d) other monosyllable. We then compare how these types match a predefined set of possible metrical patterns (i.e. sequences of strong and weak line positions of the same length). For this purpose, we propose a rule-based algorithm based on the generative rules of Czech verse published by Miroslav Červenka (2006). These rules were originally developed to capture the most common rhythmic configurations in particular meters (metrical lines) while excluding less common ones (unmetrical lines). However, as Červenka himself often observes, "metricality" itself is more of a continuous variable than a Boolean one. We therefore propose a weight called the "metrical index" meant to quantify the extent to which particular lines fulfill the requirements of particular meters. This approach proves to be reliable enough to distinguish particular meters.

Finally we discuss the ambiguities of traditional prosodic terminology. Here we argue that just like a single line, an entire poem may also fulfill the requirements of two or more different meters to the same extent (the same degree of metricality). On this basis, we reformulate the meter recognition problem as a multi-label classification task.

Czech Verse Processing System KVĚTA—Phonetic and Metrical Components

Petr Plecháč

ICL, Czech Academy of Sciences, Prague, Czech Republic

Glottotheory 7.2, 2016 (De Gruyter)

The paper describes the algorithms of the phonetic and metrical components of the Czech verse processing system KVĚTA. As such it updates the information contained in previous reports (Ibrahim et al. 2011, 2014; Plecháč and Ibrahim 2013a,b). This system is currently being used to build the Corpus of Czech Verse (CCV), which so far contains 1,689 Czech books of poetry (over 2.5 million lines) from the 19th and early 20th centuries.

In contrast to standard language corpora, each lexical unit in CCV features not only the assigned lemma and morphological tag attributes but also a phonetic transcription. Furthermore, each verse line receives the following attributes: metre (iamb, trochee...), length (number of feet), ending (feminine, masculine...), and metrical pattern. At higher levels, rhyme pairs (or n-tuples) and fixed forms (sonnet, rondel, etc.) are also annotated.

We focus here especially on the components that provide phonetic and metrical annotation: **(1) the F-component** has the task of deriving a phonetic transcription from the input data. Unlike in previous versions of the system, the algorithm in this case relies on morphologically annotated data (as provided by the stochastic tagger MorphoDiTa), which enables not only to classify monosyllabic words according to the probability of their bearing stress, but also to distinguish between stress-attracting prepositions and their homonyms (*se*: preposition/pronoun, *bez, při*: preposition/noun). Beside common rules of Czech phonetic transcription, we implement several phonotactic rules proposed by Bičan (2013); human interaction is required in the case of ambiguous bigrams (*au*, *ou*, *eu*) and loanwords (this requirement is triggered whenever a sequence occurs that has a low phonotactic probability in Czech). Based on random samples, we assess the accuracy of phonetic transcription in CCV to be as high as 0.9991 (i.e. approximately nine of every 10,000 sounds are incorrect, however none of the errors found causes a syllable miscount).

The task of **(2) the G-component** is to generate the set of all possible metrical interpretations of the input data, i.e. all the potential patterns consisting of strong and weak positions. Beside the common accentual-syllabic meters (iamb, trochee, dactyl...), several accentual-syllabic imitations of quantitative meters (e.g. dactylic hexameter and Sapphic stanza) are also taken into account.

Finally, the task of **(3) M-component** is to select the most likely metrical interpretation. For this purpose, we employ a simple Naive Bayes classifier that was trained on CCV (i.e. annotated by a previous version of the system and thoroughly manually corrected). Based on our estimation, the accuracy of the M-component surpasses 0.95 (in terms of correctly recognized lines) and reaches almost 0.98 when polymetric poems are excluded.

A Collocation-Driven Method of Discovering Rhymes (in Czech, English, and French Poetry)

Petr Plecháč

ICL, Czech Academy of Sciences, Prague, Czech Republic

Taming the Corpus. From Inflection and Lexis to Interpretation, 2017 (Springer)

The paper proposes an unsupervised language-independent method of discovering rhymes in a corpus of poetic texts. I argue that any rule-based algorithm for rhyme detection will depend greatly on the philological background of its creator since this affects the extent to which “imperfect rhymes” are considered. Moreover such an algorithm will be applicable to a certain time period only (given spelling changes). I therefore propose an algorithm which does not focus on precise sound-matching but instead relies on the probabilities of two words rhyming together derived mainly from the analyzed texts themselves.

The algorithm builds on the assumption that any sufficiently large body of poetic texts will contain recurring rhyme pairs. I employ an adaptation of the usual collocation extraction technique (t-score) to identify line-endings that co-occur more often than would be expected by chance. I show that—depending on the degree of inflection in the given language and the corpus size—up to 18% of all rhymes may be discovered in this way.

Next I use the output as a training set for simple machine learning. Each word in the training set is split into syllable peaks and consonant clusters and a Naive Bayes classifier is used to learn the probabilities of these component-pairs co-occurring in rhyme. Besides these phonetic features, I also use probabilities based on line-ending character trigrams. This is a safety net for rhyme pairs that are affected by changes to pronunciation over time (e.g. Shakespeare's *near*, original pronunciation [NE:r] – *there* [DE:r]; the idea behind this is that some other words may have preserved the original pronunciation of the given trigram and may, thus, be found in the training set; e.g. *wear*, *pear*, *swear* in this case). The trained model is then used to recognize rhymes in the corpus. The output is applied as a new training set, and a new learning phase is then completed. These iterations continue until there are no further improvements and the training set and classification output are found to be equal.

I measure the precision and recall of the algorithm against the gold standards of three corpora of poetry in different languages: Czech (~2.7 million lines), English (~93 thousand lines), and French (~27 thousand lines). With F_1 scores ranging from 0.9 to 0.95, the algorithm is shown to significantly outperform not only the rule-based detection of rhymes but also the expectation-maximization algorithm proposed in Reddy et al. 2011a.

Update (2021): As of 2019, the phonetic transcription component has been switched from MaryTTS to eSpeak. The system is now available on `PyPI` as a `RhymeTagger` package with pre-trained models for Czech, Dutch, English, French, German, Russian, and Spanish.

Versification and Authorship Attribution. A Pilot Study on Czech, German, Spanish, and English Poetry

Petr Plecháč[1], Klemens Bobenhausen[2], and Benjamin Hammerich[2]

[1]ICL, Czech Academy of Sciences, Prague, Czech Republic
[2]Project Metricalizer, Freiburg im Brsg., Germany

Studia Metrica et Poetica 5.2, 2018 (University of Tartu)

This article describes pilot experiments performed as one part of a long-term project examining the possibilities for using versification analysis to determine the authorships of poetic texts.

We first outline several reasons why versification features are suitable for authorship attribution studies. These include the Boolean nature of most of these features, their topic-independence, and their low correlation with features commonly used in stylometry. Next we briefly introduce the history of authorship attribution methods, concluding that versification features have only rarely been used in this research, and where they were involved, the analysis was based on the simple and ill-suited methods of univariate descriptive statistics.

We therefore propose an approach that combines versification analysis with multivariate models and state-of-the-art machine learning techniques. Since the audience for this study includes both stylometry experts and verse scholars, we first describe the common classifiers used in contemporary stylometry (Burrows' Delta, Argamon's Quadratic Delta, Smith–Aldridge's Cosine Delta, and Support Vector Machines). We also explain how these classifiers work and provide graph-based examples. We then evaluate their performance with versification features in Czech, German, Spanish, and English poetry. We show that authorship attribution based solely on versification features not only significantly exceeds the random baseline but in several cases also outperforms the analysis based on common stylometric features (frequencies of the most common words and character *n*-grams). Most importantly, accuracy is always greatest when all feature spaces (versification, words, *n*-grams) are concatenated into one.

Assessing the Reliability of Stress as a Feature of Authorship Attribution in Syllabic and Accentual Syllabic Verse

Petr Plecháč[1] and David J. Birnbaum[2]

[1]ICL, Czech Academy of Sciences, Prague, Czech Republic
[2]University of Pittsburgh, USA

Quantitative Approaches to Versification, 2019 (ICL, Czech Academy of Sciences)

This work builds on a recent study by one of the authors, which shows that statistics about versification may be used as a feature in the process of authorship attribution. One such statistic is what we have called the stress profile of a poem, a vector consisting of frequencies of stressed syllables at particular metrical positions.

Our initial hypothesis was that because syllabic versification (SV) regulates by definition the number of syllables in a line but not the distribution of stresses, it allows authors to individualize their rhythmical style much more than accentual syllabic versification (ASV), where the distribution of stresses is primarily determined by meter. For that reason, we expected the stress profile to be a more reliable indicator of authorship in Spanish SV than in Czech or German ASV. This hypothesis, however, was not supported by our analysis. For most of our samples, German ASV had lower accuracy than Spanish, which we had predicted, but, contrary to our expectations, the accuracy for Czech ASV and Spanish SV were more or less the same.

This result led us to hypothesize further that the traditional labels SV and ASV were misleading and we sought to measure the tonic entropy of our data. In this case, Spanish SV, as expected, was found to be the least tonically regular, while there was a significant difference between the two ASV systems: the values for Czech were even closer to Spanish than to the low-scoring German system. This explains why our initial grouping of Czech and German together into a single ASV category was insufficiently nuanced.

Relative Contributions of Shakespeare and Fletcher in Henry VIII: An Analysis Based on Most Frequent Words and Most Frequent Rhythmic Patterns

Petr Plecháč

ICL, Czech Academy of Sciences, Prague, Czech Republic

Digital Scholarship in the Humanities 36.2, 2021 (Oxford University Press)

The verse play *Henry VIII (H8)* is today widely recognized as a collaborative work and not solely the creation of William Shakespeare. However opinions differ as to the identity of the collaborator(s), with candidates including John Fletcher, Philip Massinger, and an unknown author) and the precise extent of their authorial contribution. This study aims to contribute to the question of the play's authorship using a combination of vocabulary and versification analyses and modern machine learning techniques.

As a training set, I use four plays by Shakespeare (*Coriolanus, Cymbeline, The Winter's Tale, and The Tempest*), four plays by Fletcher (*Valentinian, Monsieur Thomas, The Woman's Prize, and Bonduca*), and three plays by Massinger (*The Duke of Milan, The Unnatural Combat, and The Renegado*), all of which date from roughly the period when *H8* was supposedly written. I use a support vector machine as a classifier and the frequencies of the 500 most common rhythmic types and 500 most common words as a feature set. Cross-validation indicates that the model is highly accurate: in nine of the 11 plays, the author was recognized correctly in 100% of particular scenes, and the rate was 99% and 96% in the remaining two plays.

Classification of particular scenes of *H8* indicates that Massinger's participation is rather unlikely, while the probability of Fletcher's involvement is very high. Furthermore, the model's predictions about the share of Shakespeare's and Fletcher's respective contributions largely correspond with the canonical attributions by James Spedding (1850). The two exceptions are the second scene of Act 3, where Spedding presumed mixed authorship, and the first scene of Act 4, which he originally attributed to Fletcher.

As some studies suggest that—at least in the case of the second scene of Act 3—the shift in authorship did not happen around the start or end of the scene, I also employ a method known as rolling attribution (Eder 2016). Unlike in standard classification, in rolling attribution neither the entire text nor its logical parts (chapters, scenes etc.) are classified. Instead I focus on overlapping parts of fixed length. Cross-validation of the training set shows the accuracy of rolling attribution to be as high as 0.9977. When applied to *H8*, this method suggests that particular scenes are indeed mostly the work of a single author. This reaffirms the attributions proposed by Spedding. The main differences between my results and Spedding's attributions relate to the ambivalent outputs of my models for both scenes in Act 4. However, it is worth noting that Spedding himself expressed some doubts about the authorship of these scenes. Other differences are rather marginal and generally support the modifications proposed by Thomas Merriam (2003a,b, 2018) to Spedding's original attributions.

Beowulf Single-Authorship Claim is Unsupported

Petr Plecháč[1], Andrew Cooper[2], Benjamin Nagy[3], and Artjoms Šeļa[4,5]

[1]ICL, Czech Academy of Sciences, Prague, Czech Republic
[2]Uppsala University, Sweden
[3]University of Adelaide, Australia
[4]Institute of Polish Language, Polish Academy of Sciences, Poland
[5]University of Tartu, Estonia

Nature Human Behaviour 5, 2021 (Springer Nature)

Neidorf et al. 2019's quantitative stylometric profile of Old English verse used several novel methods. Their study supports the unitary authorship of *Beowulf* and Cynewulfian authorship of *Andreas* based on the ostensible homogeneity of a variety of features. The authors provided full access to their code and data, which demonstrates a praiseworthy commitment to open and replicable science. However, we argue that the methods presented are unsuitable for their purposes. Our replication study uses their unmodified data and text preprocessing steps, identifies errors in their analyses, questions the reliability of their results, and shows significant stylistic heterogeneity in *Beowulf*.

The authors' first argument is based on sense-pauses, identified by the presence of certain punctuation marks, all of which are editorial emendations. The unreliability of this method is demonstrated when the authors find "a marked difference in the intraline-to-total sense-pause ratio between *Genesis A* and *B*" and claim that it "can distinguish between passages of Old English verse about similar subject matter but composed by different poets" (p. 3). In fact, due to a coding error, both samples are from *Genesis A*, and the radically metrically different *Genesis B* is not included in the analysis. The "marked difference" is thus found between two pieces of a single text, the unitary authorship of which has never been questioned. Furthermore, we show that the metric is unstable even within a single text.

Next we demonstrate that the analysis of *Beowulf*'s metre is inconclusive. In particular, we note that: (1) the two halves of a line are not statistically independent and (2) the authors' slope-based methods, which use Pearson's r, are not powerful enough to reliably compare pattern distributions. We also question the slope-based analysis in case of hapax compounds and show that it is prone to both false positive and false negative results.

We were unable to replicate the results of the shared compounds analysis. Based on our re-analysis of the code and data provided by the authors, the two parts of *Beowulf* are far less correlated than is indicated by the authors' figures. Additionally, by considering other texts, we show that their method is sensitive to common topics, and thus it is expected that the two halves of *Beowulf* be correlated.

Finally, we question the results of the authors' cluster analysis based on the $m = 25$ most frequent character n-grams. We argue that this is an arbitrary choice of parameters and examine the robustness of this analysis using various alternative values of m and n. We conclude that the original value of m is too low to manage the variance in the underlying data. In contrast, when large enough values of m are used, the clustering consistently splits the text into two continuous parts, with the break occurring around the point where the scribal hand changes in the manuscript.

Summary

This thesis comprises articles on corpus verse studies and stylometry published during the last decade. Its main aim is to show how both rule-based systems and machine-learning algorithms can be used to recognize versification features (such as meter and rhyme) and, in turn, how these automatically recognized features can be used to solve authorship recognition tasks concerning poetic texts.

The first group of articles deals with problems of building an automatically annotated corpus of poetic texts. It offers two ways of recognizing poetic meter that both lead to satisfactory results in Czech poetry: (1) a rule-based system with a purely philological basis and (2) a supervised-machine learning approach. The third article offers a language-independent unsupervised method of rhyme recognition.

The second group of articles explores the options for using these automatically detected versification features for automatic authorship recognition. The first article tests the performance of versification-based authorship recognition with verse corpora in four different languages. The second article examines potential relations between recognition accuracy and versification typology. The final two articles apply versification-based authorship recognition to two cases involving disputed authorship: The first concerns the verse play *Henry VIII*, which was originally published under the name of William Shakespeare, but is believed by many to contain parts that were actually written by John Fletcher or perhaps even by other authors. The second relates to the authorship of the Old English poem *Beowulf*.

References

Alexander, Peter (1931). “Conjectural History, or Shakespeare’s Henry VIII”. In: *Essays and Studies* 16, pp. 85–119.

Argamon, Shlomo (2008). “Interpreting Burrows‘s Delta: Geometric and probabilistic foundations”. In: *Literary and Linguistic Computing* 23.2, pp. 131–147. DOI: 10.1093/llc/fqn003.

ARTFL (2009). *ARTFL: American and French research on the Treasury of the French Language*. Accessed: 2017-03-01. Centre National de la Recherche Scientifique/University of Chicago. URL: http://artfl-project.uchicago.edu/content/artfl-frantext.

Bičan, Aleš (2013). *Phonotactics of Czech*. Frankfurt am Main: Peter Lang.

Birnbaum, David J (2018). “Strong and weak metrical positions”. In: URL: http://poetry.obdurodon.org/metrical-analysis.xhtml.

Bobenhausen, Klemens (2011). “Current Trends in Metrical Analysis”. In: ed. by Christoph Küper. Frankfurt am Main: Peter Lang. Chap. The Metricalizer: Automated Metrical Markup for German Poetry, pp. 119–131.

Bobenhausen, Klemens and Günter Gehl (2009). “Jahrbuch für Computerphilologie 9”. In: ed. by Georg Braungart, Peter Gendolla, and Fotis Jannidis. Paderborn: Mentis. Chap. Automatisches metrisches Markup deutschsprachiger Gedichte, pp. 61–85.

Bobenhausen, Klemens and Benjamin Hammerich (2015). “Métrique littéraire, métrique linguistique et métrique algorithmique de l’allemand mises en jeu dans le programme Metricalizer2”. In: *Langages* 199, pp. 67–87. DOI: 10.3917/lang.199.0067.

Boyle, Robert (1886). “Henry VIII: An investigation into the origin and the authorship of the play”. In: *Transactions of the New Shakspere’s Society 1880–8*, pp. 443–487.

Burrows, John Frederick (2002). “»Delta«: a measure of stylistic difference and a guide to likely authorship”. In: *Literary and Linguistic Computing* 17.3, pp. 267–287. DOI: 10.1093/llc/17.3.267.

Burrows, John Frederick (2003). “Questions of authorship: attribution and beyond”. In: *Computers and the Humanities* 37.1, pp. 5–32. DOI: 10.1023/A:1021814530952.

Červenka, Miroslav (1965). “Nový projekt statistického rozboru verše”. In: *Česká literatura* 13, pp. 541–544.

Červenka, Miroslav (1971). *Statistické obrazy verše*. Praha: ÚČSL.

Červenka, Miroslav (2006). *Kapitoly o českém verši*. Praha: Karolinum.

Červenka, Miroslav and Květa Sgallová (1967). “On a Probabilistic Model of the Czech Verse”. In: *Prague Studies in Mathematical Linguistics 2*. Ed. by L. Doležal, P. Sgall, M. Těšitelová, and J. Vachek. Praha: Academia, pp. 105–120.

Cooper, Andrew (2017). “Unified Account of the Old English Metrical Line”. PhD thesis. Stockholm University.

Crystal, David (2007). *Original pronunciation transcriptions of Shakespeare’s Sonnets*. Accessed: 2017-03-01. URL: http://www.davidcrystal.com/books-and-articles/shakespeare.

Dobritsyn, Andrei (2016). “Rhythmic entropy as a measure of rhythmic diversity (The example of the Russian iambic tetrameter)”. In: *Studia Metrica et Poetica* 3.1, pp. 33–52. DOI: 10.12697/smp.2016.3.1.02.

Eder, Maciej (2011). “Style markers in authorship attribution. A cross-language study of the authorial fingerprint”. In: *Studies in Polish Linguistics* 6, pp. 99–114.

Eder, Maciej (2013). “Does size matter? Authorship attribution, small samples, big problem”. In: *Digital Scholarship in the Humanities* 30.2, pp. 167–182. DOI: 10.1093/llc/fqt066.

Eder, Maciej (2016). “Rolling stylometry”. In: *Digital Scholarship in the Humanities* 31.3, pp. 457–469. DOI: 10.1093/llc/fqv010.

Eder, Maciej (2017). “Short samples in authorship attribution: A new approach”. In: Montreal: McGill University, pp. 221–224. URL: https://dh2017.adho.org/abstracts/341/341.pdf.

Ege, Karl (1922). “Shakespeare’s Anteil an “Henry VIII””. In: *Shakespeare’s Jahrbuch* 58, pp. 99–115.

Eisen, Mark, Alejandro Riberio, Santiago Segarra, and Gabriel Egan (2017). “Stylometric analysis of early modern period English plays”. In: *Digital Scholarship in the Humanities* 33.3, pp. 500–528. DOI: 10.1093/llc/fqx059.

Fabb, Nigel (2015). *What is Poetry?: Language and Memory in the Poems of the World*. Cambridge: Cambridge University Press.

Al-Falahi, Ahmed, Mohamed Ramdani, and Mostafa Bellafkih (2017). “Machine Learning for Authorship Attribution in Arabic Poetry”. In: *International Journal of Future Computer and Communication* 6.2, pp. 42–46.

Farnham, Willard (1916). “Colloquial contractions in Beaumont, Fletcher, Massinger and Shakespeare as a test of authorship”. In: *Publications of the Modern Language Association of America* 31, pp. 326–358.

Fleay, Frederick Gard (1874). “On the authorship of The Taming of the Shrew”. In: *Transactions of the New Shakespeare Society* 1, pp. 85–129.

Fleay, Frederick Gard (1876). *Shakespeare Manual*. London: Macmillan.

Fleay, Frederick Gard (1885). “Mr. Boyle’s theory as to “Henry VIII””. In: *Athenæum* 2994, p. 355.

Fleay, Frederick Gard (1886). *A Chronicle History of Life and Work of William Shakespeare*. London: John C. Nimmo.

Furnivall, Frederick James (1874). “Another fresh confirmation of Mr. Spedding’s division and date of the play of Henry VIII”. In: *Transactions of the New Shakspere’s Society 1*, appendix 6–7.

Gardner, Martin (1978). “The bells: Versatile numbers that can count partitions of a set, primes and even rhymes”. In: *Scientific American* 238, pp. 24–30.

Gasparov, Mikhail (1996). *A History of European Versification*. Oxford: OUP.

Golston, Chris (2009). “Old English Feet”. In: *Versatility in Versification*. Ed. by Tonya Kim Dewey and Frog. New York, NY: Peter Lang, pp. 105–122.

Grieve, Jack (2005). “Quantitative Authorship Attribution. A History and an Evaluation of Techniques”. MA thesis. Burnaby: Simon Fraser University.

Grieve, Jack (2007). “Quantitative Authorship Attribution: An Evaluation of Techniques”. In: *Literary and Linguistic Computing* 22.3, pp. 251–270. DOI: 10.1093/llc/fqm020.

Grzybek, Peter (2007). “The emergence of stylometry: prolegomena to the history of term and concept”. In: *Text within Text – Culture* *within Culture*. Ed. by Katalin Kroó and Peeter Torop. Budapest and Tartu: L'Harmattan, pp. 251–270.

Haider, Thomas (2021). "Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features". In: *preprint*. URL: https://arxiv.org/abs/2102.08858.

Hajič, Jan (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Praha: Karolinum.

Hart, Alfred (1941). "Vocabularies of Shakespeare's plays". In: *The Review of English Studies* 19.74, pp. 128–140.

Hickson, Samuel (1850). "Who wrote Shakespeare's Henry VIII". In: *Notes and Queries* 2, p. 198.

Hoover, David L. (2004a). "Delta Prime?" In: *Literary and Linguistic Computing* 19.4, pp. 477–495. DOI: 10.1093/llc/19.4.477.

Hoover, David L. (2004b). "Testing Burrows's Delta". In: *Literary and Linguistic Computing* 19.4, pp. 453–475. DOI: 10.1093/llc/19.4.453.

Horton, Thomas Bolton (1987). "The Effectiveness of the Stylometry of Function Words in Discriminating between Shakespeare and Fletcher". PhD thesis. University of Edinburgh.

Hoy, Cyrus (1962). "The shares of Fletcher and his collaborators in the Beaumont and Fletcher Canon VII". In: *Studies in Bibliography* 15, pp. 71–90.

Ibrahim, Robert and Petr Plecháč (2011). "Toward Automatic Analysis of Czech Verse". In: *Formal Approaches in Poetics*. Ed. by B. P. Scherr, J. Bailey, and E. V. Kazartsev. Lüdenscheid: RAM, pp. 295–305.

Ibrahim, Robert and Petr Plecháč (2014). "La Théorie du vers et le Cercle linguistique du Prague". In: *La Linguistique* 50.2, pp. 101–114. DOI: 10.3917/ling.502.0101.

Ingram, John Kells (1874). "On the "weak endings" of Shakespeare, with some account of the history of the verse tests in general". In: *Transactions of the New Shakespeare Society* 1, pp. 442–426.

Jackson, MacDonnald P. (1997). "Phrase lengths in Henry VIII: Shakespeare and Fletcher". In: *Notes and Queries* 44.1, pp. 75–80.

Jakobson, Roman (1995). "Základy českého verše". In: *Základy českého verše*. Ed. by M. Červenka. Jinočany: H&H, pp. 157–248.

Jannidis, Fotis, Steffen Pielström, Christof Schöch, and Thorsten Vitt (2015). "Improving Burrows' Delta. An empirical evaluation of text distance measures". In: *Digital Humanities 2015: Conference abstracts*. URL: http://dh2015.org/abstracts/.

Jelínek, Tomáš and Vladimír Petkevič (2011). "Systém jazykového značkování korpusů současné psané češtiny". In: *Korpusová lingivistika Praha – 2011: Gramatika a značkování korpusů*. Ed. by V. Petkevič and Rosen. A. Praha: NLN, pp. 154–170.

Juola, Patrick (2006). "Authorship Attribution". In: *Foundations and Trends in Informational Retrieval* 1.3, pp. 233–334. DOI: 10.1561/1500000005.

Kendall, Calvin B. (1991). *The Metrical Grammar of Beowulf*. Cambridge: Cambridge University Press.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon (2009). "Computational methods in authorship attribution". In: *Journal of the American Society for Information Science and Technology* 60.1, pp. 9–26. DOI: 10.1002/asi.20961.

Levý, Jiří (1964a). "Matematický a experimentální rozbor verše". In: *Česká literatura* 12, pp. 181–213.

Levý, Jiří (1964b). "Předběžné poznámkyy k informační analýze verše". In: *Slovenská literatúra* 9, pp. 15–37.

Levý, Jiří (1971). "Matematické aspekty teorie verše". In: *Bude literární věda exaktní vědou?* Ed. by L. Doležal, P. Sgall, M. Těšitelová, and J. Vachek. Praha: ČS, pp. 264–288.

Levý, Jiří (2011). *The Art of Translation*. Amsterdam.

Lotman, Mihhail (2015). "A study on Shakespeare's verse in its historical context (Marina Tarlinskaja, Shakespeare and the Versification of English Drama, 1561–1642, Ashgate, 2014)". In: *Studia Metrica et Poetica* 2.1, pp. 140–153. DOI: 10.1093/llc/fqt066.

Lotman, Yuri and Mihhail Lotman (1986). "Pushkin: Issledovanija i materialy XII". In: ed. by Nina Nikolaevna Petrunina. Leningrad: Nauka. Chap. Vokrug desjatoj glavy "Evgenija Onegina", pp. 124–151.

Malone, Edmond (1787). *A dissertation on parts one, two and three of Henry the Sixth tending to shew that those plays were not written originally by Shakespeare*. London: Henry, Baldwin.

MaryTTS (2017). *MaryTTS: An open source, multilingual text-to-speech synthesis system*. Accessed: 2017-03-01. URL: http://github.com/marytts/marytts.

Maxwell, Baldwin (1923). “Fletcher and Henry the Eighth”. In: *The Manly Anniversary Studies in Language and Literature*. Chicago: University of Chicago, pp. 104–112.

Mendenhall, Thomas Corwin (1887). “The characteristic curves of composition”. In: *Science* 9, pp. 237–249.

Mendenhall, Thomas Corwin (1901). “A mechanical solution to a literary problem”. In: *Popular Science Monthly* 60, pp. 97–105.

Merriam, Thomas (1979). “What Shakespeare wrote in “Henry VIII”: part one”. In: *The Bard* 2.3, pp. 81–94.

Merriam, Thomas (1980). “What Shakespeare wrote in “Henry VIII”: part two”. In: *The Bard* 2.4, pp. 111–118.

Merriam, Thomas (2003a). “Taylor’s method applied to Shakespeare and Fletcher”. In: *Notes and Queries* 50.4, pp. 419–423.

Merriam, Thomas (2003b). “Though this be suplementarity, yet there is method in’t”. In: *Notes and Queries* 50.4, pp. 423–426.

Merriam, Thomas (2018). “Henry VIII, All is True?” In: *Notes and Queries* 65.1, pp. 84–88.

Mikros, George and Kostas Perifanos (2013). “Authorship Attribution in Greek Tweets Using Author’s Multilevel N-Gram Profiles”. In: *Papers from the 2013 AAAI Spring Symposium. “Analyzing Microtext”, 25–27 March 2013, Stanford, California*. Palo Alto: AAAI Press, pp. 17–23. URL: https://www.aaai.org/ocs/index.php/SSS/SSS13/paper/view/5714.

Mosteller, Frederick and David Wallace (1964). *Inference and Disputed Authorship: The Federalist*. Reading: Addison-Wesley.

Navarro-Colorado, Borja (2015). “A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects”. In: *Computational Linguistics for Literature NAACL 2015*. Denver. URL: http://www.aclweb.org/anthology/W/W15/W15-0712.pdf.

Navarro-Colorado, Borja, María Ribes-Lafoz, and Noelia Sánchez (2016). “Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož. URL: http://www.aclweb.org/anthology/W/W15/W15-0712.pdf.

Neidorf, Leonard, Madison S. Krieger, Michelle Yakubek, Pramit Chaudhuri, and Joseph P. Dexter (2019). “Large-scale quantitative profiling of the Old English verse tradition”. In: *Nature*

Human Behaviour 3.6, pp. 560–567. DOI: 10.1038/s41562-019-0570-1.

Oliphant, Ernest Henry (1891). “The works of Beaumont and Fletcher”. In: *Englische Studien* 15, pp. 321–360.

Oras, Ants (1953). “»Extra monosyllables« in Henry VIII and the problem of authorship”. In: *Journal of English and Germanic Philology* 52, pp. 198–213.

Palková, Zdena (1994). *Fonologie a fonetika češtiny*. Praha: Karolinum.

Piera, Carlos José (1980). “Spanish Verse and the Theory of Meter”. PhD thesis. Los Angeles: University of California.

Pilschchikov, Igor and Anatoli Starostin (2011). “Automated Analysis of Poetic Texts and the Problem of Verse Meter”. In: *Current Trends in Metrical Analysis*. Ed. by C. Küper. Bern et al.: Peter Lang, pp. 133–140.

Platt, John (1999). “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods”. In: *Advances in Large Margin Classifiers* 10.3, pp. 61–74.

Plecháč, Petr (2008). “Česká versifikace a generativní metrika”. In: *Aluze* 11, pp. 86–93.

Plecháč, Petr (2016). “Czech Verse Processing System KVĚTA – Phonetic and Metrical Components”. In: *Glottotheory* 7.2, pp. 159–174. DOI: 10.1515/glot-2016-0013.

Plecháč, Petr (2018). “Taming the Corpus. From Inflection and Lexis to Interpretation”. In: ed. by Masako Fidler and Václav Cvrček. Cham: Springer. Chap. A collocation-driven method of discovering rhymes (in Czech, English, and French poetry), pp. 79–95.

Plecháč, Petr and David J Birnbaum (2019). “Assessing the reliability of stress as a feature of authorship attribution in syllabic and accentual syllabic verse”. In: *Quantitative Approaches to Versification*. Ed. by Petr Plecháč, Barry P Scherr, Tatyana Skulacheva, Helena Bermúdez-Sabel, and Robert Kolár. ICL, pp. 13–18. URL: http://versologie.cz/conference2019/proceedings.php.

Plecháč, Petr, Klemens Bobenhausen, and Benjamin Hammerich (2018). “Versification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry”. In: *Studia Metrica et Poetica* 5.2, pp. 29–54. DOI: 10.12697/smp.2018.5.2.02.

Plecháč, Petr and Robert Ibrahim (2013a). “Automatyczna analiza wiersza: punkt wyjściay”. In: *Potencjał wiersza*. Ed. by W. Sadowski. Warszawa: IBL, pp. 273–280.

Plecháč, Petr and Robert Ibrahim (2013b). “Phonological and Morphological Means Compensating for Non-Metricality in 19th Century Czech Verse”. In: *Prace Filologiczne* 59.3, pp. 31–50.

Plecháč, Petr and Robert Kolár (2015). “The Corpus of Czech Verse”. In: *Studia Metrica et Poetica* 2.1, pp. 107–118. DOI: 10.12697/smp.2015.2.1.05.

Reddy, Sravana and Kevin Knight (2011a). “Unsupervised discovery of rhyme schemes”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland: ACL, pp. 77–82. URL: https://www.aclweb.org/anthology/P11-2014/.

Reddy, Sravana and Kevin Knight (2011b). *Unsupervised discovery of rhyme schemes. The code.* Accessed: 2017-03-01. URL: https://github.com/sravanareddy/rhymediscovery.

Roderick, Richard (1758). *Remarks on Shakespeare*. Reprinted in William Shakespeare: The Critical Heritage 4 (1976).

Savoy, Jacques (2020). *Machine Learning Methods for Stylometry - Authorship Attribution and Author Profiling*. Springer. DOI: 10.1007/978-3-030-53360-1.

Segarra, Santiago, Mark Eisen, and Alejandro Riberio (2013). “Authorship attribution using function words adjacency networks”. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5563–5567. DOI: 10.1109/ICASSP.2013.6638728.

Sgallová, Květa (1964). “Vyyužití moderní technikyy při rozboru verše”. In: *Česká literatura* 12, pp. 158–165.

Sgallová, Květa (1999). “Thesaurus českých meter”. In: *Česká literatura* 47, pp. 286–289.

Shapir, Maxim (1997). “Fenomen Batenkova i problema mistifikatsii (lingvostikhovedcheskij aspekt 1–2”. In: *Philologica* 4, pp. 85–144.

Shapir, Maxim (1998). “Fenomen Batenkova i problema mistifikatsii (lingvostikhovedcheskij aspekt 3–5”. In: *Philologica* 5, pp. 49–132.

Sievers, Eduard (1983). *Altgermanische Metrik*. Halle: Niemeyer.

Simpson, Edward (1949). “Measurement of diversity”. In: *Nature* 163, p. 688.

Skarnitzl, Radek (2014). “O slovním přízvuku na jednoslabičných předložkách v češtině”. In: *Naše řeč* 97.2, pp. 78–91.

Smith, P. W. H. and W. Aldridge (2011). “Improving authorship attribution: Optimizing Burrows’ Delta method”. In: *Journal of Quantitative Linguistics* 18.1, pp. 63–88. DOI: 10.1080/09296174.2011.533591.

Sonderegger, Morgan (2011). “Applications of graph theory to an English rhyming corpus”. In: *Computer Speech and Language* 25, pp. 655–678. DOI: 10.1016/j.csl.2010.05.005.

Spedding, James (1850). “Who wrote Shakespeare’s Henry VIII”. In: *The Gentleman’s Magazine*, pp. 115–123.

Stamatatos, Efstathios (2009). “A Survey of Modern Authorship Attribution Methods”. In: *Journal of the Association for Information Science and Technology* 60.3, pp. 538–556. DOI: 10.1002/asi.21001.

Stockwell, Robert P. and Donka Minkova (1997). “Prosody”. In: *A Beowulf Handbook*. Ed. by Robert E. Bjork and John D. Niles. Lincoln, NE: University of Nebraska Press, pp. 55–85.

Štraus, František (2002). *Základy informačnej analýzy verše*. Bratislava: UK.

Sykes, Henry Dugdale (1919). *Sidelights on Shakespeare*. Stratford-upon-Avon: The Shakespeare Head Press.

Tarlinskaja, Marina (1987). *Shakespeare’s Verse: Iambic Pentameter and the Poet’s Idiosyncrasies*. New York.

Tarlinskaja, Marina (2014). *Shakespeare and the versification of English Drama, 1561– 1642*. Farnham: Ashgate.

Tomashevsky, Boris Viktorovich (2008). “Izbrannye raboty o stikhe”. In: Moskva and Sankt-Peterburg: Akademija. Chap. Pjatistopnyj jamb Pushkina, pp. 201–213.

Vickers, Brian (1976). *William Shakespeare: The Critical Heritage 4*. London and New York: Routledge.

Vickers, Brian (2004). *Shakespeare, Co-Author: A Historical Study of the Five Collaborative Plays*. Oxford: Oxford University Press.

Weber, Henry (1812). *The Works of Beaumont and Fletcher in Fourteen Volumes. Vol. 13*. Edinburgh: J. Ballantyne & Co.

Williams, Carrington Bonsor (1975). “Mendenhall’s studies of word-length distribution in the works of Shakespeare and Bacon”. In: *Biometrika* 62.1, pp. 207–212. DOI: 10.1093/biomet/62.1.207.

Yule, George Udny (1938). “On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship”. In: *Biometrika* 30, pp. 363–390.

Zabaleta, Manex Aguirrezabal (2017). “Automatic Scansion of Poetry”. PhD thesis. Bilbao: Euskal Herriko Unibertsitatea.