Univerzita Karlova
Filozofická fakulta
Ústav Českého národního korpusu

Petr Plecháč

ATRIBUCE AUTORSTVÍ BÁSNICKÝCH TEXTŮ

AUTHORSHIP ATTRIBUTION OF POETIC TEXTS

Dizertační práce

Praha 2019

vedoucí dizertační práce: doc. Mgr. Václav Cvrček, Ph.D.

studijní program: Filologie

studijní obor: Matematická lingvistika

Prohlašuji, že jsem dizertační práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 11. 7. 2019 Petr Plecháč

Je mojí milou povinností poděkovat **Václavu Cvrčkovi**, jehož zájem o moji práci dalece přesahuje povinnosti plynoucí z jeho úlohy školitele a bez jehož pomoci by tato práce nikdy nemohla vzniknout. Velký dík pak patří následujícím kolegům a kamarádům, kteří mi pro potřeby této práce laskavě poskytli své dlouhá léta budované veršové korpusy, případně se mnou dlouhá léta jeden takový korpus budovali, pomáhali mi s nastavením jimi vytvořených softwarových modulů a ochotně se mnou trávili dlouhé hodiny v diskuzích, třebas i napříč několika časovými pásmy: **Robert Kolár** (Ústav pro českou literaturu AV ČR, v. v. i.), **Klemens Bobenhausen** (Metricalizer, Freiburg im Breisgau), **Benjamin Hammerich** (Metricalizer, Freiburg im Breisgau), **David Birnbaum** (University of Pittsburgh), **Artjoms Šela** (Tartu Ülikool), **Borja Navarro-Colorado** (Universidad de Alicante), **Ryan Heuser** (Stanford University), **Arthur M. Jacobs** (Freie Universität Berlin). Největší dík pak patří mé **manželce a dceři** za to, že několik let vydržely poslouchat dlouhé monology o rozpoznávání autorství a ještě mě v tom podporovaly.

ABSTRAKT

Pro rozpoznávání autorství básnických textů nabízí současná stylometrie řadu metod založených na analýze pestré škály textových rysů (např. frekvence slov, frekvence znakových *n*-gramů). Jeden podstatný aspekt těchto textů ovšem zůstává stranou, a to jejich stránka versologická. Tato práce proto na čtyřech korpusech básnických textů (českých, německých, španělských a anglických) analyzuje, do jaké míry lze versologické charakteristiky – jako např. četnosti rytmických konfigurací nebo četnosti různých typů rýmů – využít jako indikátor autorství básnického textu. Ukazujeme, že (1) úspěšnost versologických modelů vysoce převyšuje hranici *random baseline*, (2) ojediněle převyšuje úspěšnost obvyklých lexikálních modelů a (3) kombinované versologicko-lexikální modely vykazují téměř vždy vyšší úspěšnost než jednotlivé modely samy o sobě. V další části práce jsou versologické rysy využity pro atribuci dvou textů se sporným autorstvím: (1) veršované drama *The Famous History of the Life of King Henry the Eigth* poprvé otištěné pod jménem Williama Shakespeara, u nějž se ovšem před-pokládá i autorská účast Johna Fletchera, příp. dalších autorů a (2) básně publikované pod jménem Josefa Baráka, u nichž existuje hypotéza, že jejich skutečným autorem je Jan Neruda.

KLÍČOVÁ SLOVA

atribuce autorství, stylometrie, versologie, strojové učení, korpusová lingvistika

---

ABSTRACT

Contemporary stylometry offers a number of methods for authorship recognition of poetic texts based on a variety of textual features (e.g. word frequencies, frequencies of character *n*-grams). However, it seems that one important aspect of these texts has been rather left aside – this aspect is versification. The thesis uses four corpora of poetic texts (Czech, German, Spanish, and English) in order to analyze to what extent versification features – such as frequencies of rhythmic patterns or frequencies of various types of rhymes – may be used as an indicator of authorship. We show that (1) versification-based models significantly outperform the *random baseline*, (2) in some cases versification-based models even outperform the traditionally used lexical models, (3) in most of the cases combination of both types of models outperforms the given models alone. Versification features are consequently employed for the purpose of attribution of two texts of doubted authorship: (1) the versified play *The Famous History of the Life of King Henry the Eigth* which was originally published under the name of William Shakespeare, but where many suppose that some parts were actually written by John Fletcher or even other authors, and (2) poems published under the name of Josef Barák where there is a hypothesis that their real author is Jan Neruda.

KEYWORDS

authorship attribution, stylometry, versification, machine learning, corpus linguistics

Obsah

Úvod

Současná stylometrie, konkrétně ta její část zabývající se atribucí autorství, zažívá nebývalý rozmach. V posledních letech byla navržena, testována a aplikována celá řada textových charakteristik, na jejichž základě lze s pomocí metod multidimenzionální statistiky a/nebo strojového učení s velmi vysokou úspěšností odlišit texty psané různými autory. Ve většině případů se ovšem – z pohledu statistiky – jedná o tzv. řídké jevy nebo o jevy s distribucí LNRE – *large number of rare events* (např. četnosti jednotlivých slovních tvarů, slovních *n*-gramů, znakových *n*-gramů). Pro úspěšnou analýzu jsou tak obvykle potřeba poměrně rozsáhlé vzorky (tisíce až desetitisíce slov). V oblasti uměleckých textů se proto většina studií omezuje na rozsáhlé prozaické texty. Básnické texty – přestože neznámé nebo zpochybněné autorství provází nejčastěji právě je – zůstávají stranou.

Na druhou stranu, v poezii se lze opřít také o celou řadu binárních proměnných nebo proměnných, které mohou nabývat jen relativně malý počet možných hodnot, z oblasti versologie. Určitá část versologických charakteristik textu sice spadá do sféry vědomě volených a tudíž pro rozlišení autorství nevhodných – např. typ básnického metra (jamb, trochej, daktyl…), délka strofy, nebo typ rýmového schématu – u velké části z nich lze ale vědomou kontrolu předpokládat jen stěží (např. tendence při konkrétním výběru možných rytmických realizací metrického schématu, četnosti výskytu jednotlivých hlásek v rýmech). Přestože tyto charakteristiky bývají tradičně považovány za specifický rys autorské, případně alespoň dobové poetiky, při atribuci autorství byly dosud využívány pouze sporadicky.

Cílem této dizertační práce je testovat využitelnost takových versologických rysů při atribuci autorství na materiálu česky, německy, španělsky a anglicky psané poezie a aplikovat takový přístup na konkrétní případy sporného autorství.

Kapitola 1 podává stručný přehled vývoje kvantitativních metod určování autorství a detailně popisuje metody využívané v této práci.

Kapitola 2 popisuje vybrané versologické rysy a diskutuje jejich využitelnost pro potřeby atribuce autorství.

Kapitola 3 popisuje výsledky testování úspěšnosti modelů založených na versologických charakteristikách a jejich srovnání s modely obvykle využívanými ve stylometrii.

Kapitola 4 se zabývá dvěma případy sporného autorství, konkrétně: (1) veršovaným dramatem *The Famous History of the Life of King Henry the Eight*, kde existuje poměrně silná hypotéza, že jejím jediným autorem není William Shakespeare, ale že obsahuje i části napsané Johnem Fletcherem, případně i jinými autory, (2) básněmi publikovanými pod jménem Josefa Baráka, kde existuje hypotéza, že jejich skutečným autorem je Jan Neruda. V obou případech jsou při atribuci využity (mimo jiné) versologické rysy.

***

Kapitola 1 je rozšířenou verzí oddílů *(2) Motivations* a *(3) History and related works* článku *Versification and authorship attribution. Pilot study on Czech, German, Spanish, and English poetry* (Plecháč–Bobenhausen–Hammerich 2018).

Kapitola 3.2 vychází částečně z textu *A collocation-driven method of discovering rhymes (in Czech, English, and French poetry)* (Plecháč 2018).

Kapitola 4.2 je přepracovanou verzí oddílů *(2) Atribuce „Kříže pod Petřínem“ provedená Pavlem Vašákem, (3) Kvadratická kosinova Delta* a *(4) Bootstrapovaná kosinová Delta* článku *Problém Barák–Neruda z pohledu současné stylometrie* (Plecháč–Flaišman 2017).

Kopie prohlášení spoluautorů jsou součástí dizertační práce.

1 Kvantitativní metody určování autorství

Následující kapitola podává přehled vývoje stylometrie, přičemž se omezuje na oblast zabývající se atribucí autorství. Vzhledem k tomu, že historie těchto metod byla podrobně zmapována v řadě přehledových publikací (zejm. Koppel–Schler–Argamon 2009; Stamatatos 2009; Juola 2006; Grieve 2005; Holmes 1998; 1994), jsou zde pouze stručně popsány hlavní milníky. Vedle toho jsou podrobněji rozvedeny:

(1) metody využívané v této práci (kap. 1.3, 1.4);

(2) práce zabývající se versologickými rysy (kap. 1.1, 1.5), které obvykle v přehledových publikacích zmiňovány nebývají;

(3) další práce, které bývají v přehledových publikacích přehlíženy (kap. 1.3.3).

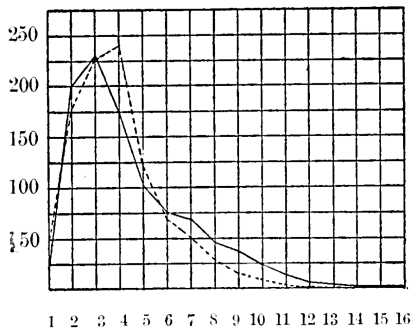
1.1 Počátky stylometrie

Většina přehledových publikací označuje za první teoretický impulz ke vzniku stylometrie coby vědecké disciplíny několik pasáží v dopisu Augusta de Morgana, za skutečného průkopníka stylometrie pak Thomase Corwina Mendenhalla, případně Williama Benjamina Smithe nebo Lucia Adelna Shermana.

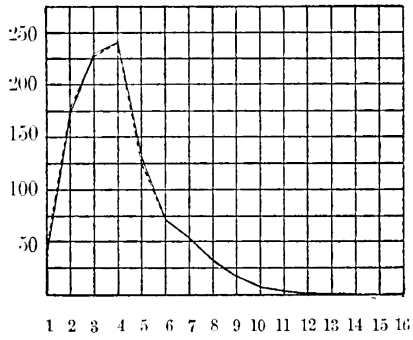
Ve zmiňovaném dopisu adresovaném reverendu Healdovi se britský matematik Augustus de Morgan (1851/1882) zamýšlí nad možnostmi určení autorství novozákonních epištol připisovaných svatému Pavlovi a uvádí kvantitativní charakteristiky textu coby možný indikátor: navrhuje, že by bylo možné odlišit texty skutečně psané Pavlem od ostatních na základě průměrné délky slov měřené počtem znaků: „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). S povzdechem pak dodává, že tato metoda by měla být testována i na jiných textech psaných v různých jazycích: „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" (1851/1882: 216).

Finance na takový experiment se ovšem podařilo sehnat až o padesát let později americkému fyzikovi Thomasi Corwinu Mendenhallovi. Ten nejprve v článku *The characteristic curve of composition* (1887) navrhnul pracovat namísto průměru s celou distribucí četností slov různé délky a tuto metodu pak díky podpoře mecenáše Augusta Hemenwaye později použil při řešení skutečné otázky sporného autorství, jehož výsledky shrnul v článku *A mechanical solution to a literary problem* (1901). Mendenhall zde porovnává tvar křivky určené relativními četnostmi slov různé délky v textech při-

pisovaných Williamu Shakespearovi s křivkami extrahovanými z textů Francise Bacona a Christophera Marlowa (viz OBR. 1.1, OBR. 1.2) a na základě jejich odlišnosti/podobnosti opatrně dovozuje, že Shakespearovy texty nemohl napsat Bacon, zatímco je velmi pravděpodobné, že jejich skutečným autorem je Marlow (1901: 104–105). Později se ovšem ukázalo, že celý experiment byl zkreslen jednou významnou vnější proměnnou, a to tím, že u Shakespeara a Marlowa zkoumal Mendenhall veršované texty zatímco u Bacona texty neveršované (srov. Williams 1975).



OBR. 1.1: Relativní četnosti (v promile) slovních délek měřených počtem znaků v textech připisovaných W. Shakespearovi (přerušovaná čára) a v textech F. Bacona (plná čára). Zdroj: Mendenhall 1901: 104 (faksimile).



OBR. 1.2: Relativní četnosti (v promile) slovních délek měřených počtem znaků v textech připisovaných W. Shakespearovi (přerušovaná čára) a v textech C. Marlowa (plná čára téměř překrývající přerušovanou). Zdroj: Mendenhall 1901: 104 (faksimile).

Nezávisle na Mendenhallovi se v 80. letech 19. století věnoval možnostem využití kvantitativních metod při atribuci autorství i americký matematik William Benjamin Smith. V článku *Curves of pauline and pseudo-pauline style* (1888a; 1888b) publikovaném pod pseudonymem Conrad Mascol se stejně jako de Morgan zabýval autorstvím Pavlových epištol a stejně jako Mendenhall použil coby rozlišovací kritérium tvar křivek reprezentujících různé vlastnosti textu (např. průměrný počet vět na stránce, průměrný počet předložek na stránce aj.). Na základě srovnání těchto ukazatelů v epištolách, u nichž panuje obecná shoda na Pavlově autorství (Řím, 1 Kor, 2 Kor, Gal), s epištolami, u nichž bylo Pavlovo autorství zpochybněno (Ef, Fp, Kol), pak konstatoval, že autor první skupiny textů pravděpodobně není totožný s autorem druhé skupiny. Za pozornost stojí, že Smith explicitně uvádí, že kritériem pro výběr textových rysů by měla být jejich co možná minimální závislost na tématu textu,[1] což je předpoklad, který – jak uvidíme později (kap. 1.2) – nebyl ještě ani ve 20. století samozřejmostí.

Jako třetí průkopnická práce, která patrně rovněž vznikla nezávisle na Mendenhallovi, bývá uváděn článek Lucia Adelna Shermana (1888) zabývající se průměrnou délkou věty měřenou počtem slov v dílech různých anglicky píšících prozaiků.[2] Možnosti využití této metriky pro atribuci autorství Sherman nezmiňuje.

1 „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) ovšem uvádí, že Shermana mohla motivovat reakce na první Mendenhallův článek otištěná ještě v roce 1887 v časopisu *Science*, kde lze mimo jiné číst: „There are other characteristics of writers equally susceptible of treatment by the statistical and graphical

Zcela ojediněle (Grzybek 2014; Grieve 2007) bývá ovšem v přehledových publikacích zmiňována jiná odnož stylometrie, jejíž počátky lze vysledovat plných sto let před Mendenhallovým prvním článkem a více než šedesát let před de Morganovým dopisem: řada atribucí provedená shakespearology na základě kvantifikace verše a rýmu.

Pravděpodobně nejstarší doklad takového přístupu lze nalézt ve studii Edmonda Malona (1787/1803), v níž je formulována hypotéza, že žádnou ze tří částí hry *Henry VI.* nenapsal ve skutečnosti Shakespeare. Malonovy argumenty jsou mimo jiné kvantitativně-versologické: opírá se o zjištění, že tyto texty obsahují mnohem méně rýmů a mnohem více případů shody syntaktického a veršového členění než ostatní Shakespearovy hry.

Dalším příklad je komentář Henryho Webera ke hře *The Two Noble Kinsmen* (1812), která byla poprvé vydána roku 1634 jako společné dílo Williama Shakespeara a Johna Fletchera. Weber atribuuje jednotlivé části konkrétnímu autorovi na základě frekvencí typů zakončení verše:

> „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).

O několik desítek let později pak na základě stejné metody formuloval James Spedding (1850) hypotézu o společném autorství Shakespeara a Fletchera u hry *Henry VIII*; podrobněji viz kap. 4.1.

K největšímu rozmachu versologicky orientované stylometrie došlo v 70. letech 19. století v rámci *New Shakspere Society*,[3] jejíž členové navrhli celou řadu „veršových testů“ (verse tests) sloužících k odlišení skutečných Shakespearových děl od textů napsaných jinými autory. V prvním ročníku *Transactions of the New Shakspere Society* tak například John Kells Ingram zavedl rozlišení nepřízvučných zakončení verše na „light endings“ a „weak endings“[4] a na základě statistiky jejich frekvencí v celém Shakespearově díle podpořil Speddingovu hypotézu o autorství hry *Henry VIII* (Ingram 1874). Sám Ingram tuto metodu pojmenoval „weak-ending test“.

Podobných testů byla navržena (případně ze starších prací přejata) a aplikována celá řada: „rhyme-test“ (srovnání četností rýmovaných veršů), „stopt-line test“ (srovnání

---

method, in which their personal pecularities 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 pecularity“ (Eddy 1887: 297).

3 „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. […] Though it has hitherto been too much to ask people to suppose that SHAKSPERE knew how to spell his own name, I hope the demand may not prove too great for the imagination of the Members of the New Society“ (Furnivall 1874a: 6).

4 „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 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).

četností shody veršového a syntaktického členění), „middle-syllable test“ (četnost extrametrických slabik na konci prvního půlverše), „caesura-test“ (srovnání četností mezislovních předělů po šesté slabice alexandrinů).[5]

Velká část atribucí provedených členy *New Shakspere Society* se sice později ukázala jako chybná (ať už kvůli jednoduchosti a mechanickému aplikování daných metod, nebo kvůli množství chyb ve zdrojových datech; srov. Grieve 2005: 6), nic to ale nemění na skutečnosti, že se jedná o neprávem opomíjenou počáteční kapitolu dějin stylometrie.

1.2 Hledání „zlatého rysu“

K dalšímu rozvoji stylometrie ve 20. století nepřímo přispěly práce George Kingsleyho Zipfa. Formulace prvního Zipfova zákona (1932), podle nějž všechny texty psané přirozeným jazykem vykazují tutéž rankovou frekvenční distribuci slov, vedla patrně k znovunastolení otázky, jestli by nebylo možné najít nějakou podobnou vlastnost, která zůstává v textech produkovaných jedním autorem stabilní, ale mění se napříč texty různých autorů (srov. Koppel–Schler–Argamon 2009: 4–5).

Mezi nejvlivnější patří v tomto období práce George Udny Yula. Jako vhodný rys pro odlišení autorství navrhoval nejprve Yule (1939) délku věty měřenou počtem slov. Na rozdíl od Shermana (viz 1.1) ovšem neporovnával jen průměrné hodnoty, ale i další charakteristiky distribuce: medián, kvartily $Q_{0,25}$, $Q_{0,75}$, mezikvartilové rozpětí a – vzhledem k tomu, že délky vět obecně vykazují silně pozitivně zešikmené log-normální rozdělení – decil $Q_{0,9}$.

Později v knize *The Statistical Study of Literary Vocabulary* (Yule 1944) navrhl pro potřeby určování autorství známou – pravděpodobně první – metriku bohatosti slovníku:

$$K = \frac{10^4[(\sum_{m=1}^{m_{max}} m^2 V_m) - N]}{N^2} \quad (1.1),$$

kde N značí délku textu měřenou počtem tokenů a V_m počet typů s absolutní frekvencí m.

Dodejme, že při konkrétních atribucích založených na této metrice se Yule rozhodl pracovat pouze se substantivy, která podle něj nejlépe charakterizují autorský styl.

> „My object in limiting myself to nouns for the investigation into the vocabularies of Thomas à Kempis and Gerson was in part simply the limitation of material and the exclusion of words of little or no significance as regards style, such as prepositions, pronouns, etc. Of the three principal parts of speech, nouns, adjectives and verbs, I thought nouns would probably be the most significant or characteristic“ (Yule 1944: 21).

Přesvědčení, že vysoce frekventovaná synsémantika nemohou nijak přispět k rozpoznání autorství, bylo v té době ostatně poměrně rozšířené (srov. Grieve 2005: 32–34).

Podobných jednoduchých charakteristik byla pro potřeby atribuce navržena celá řada – např. průměrná délka slova měřená počtem slabik (Fucks 1952) nebo četnost přejatých slov (Herdan 1956). Žádná z nich se ale neukázala být dostatečně robustní a při aplikaci na jiné atribuční úlohy než ty, pro něž byly původně navrženy, tyto metody obvykle selhaly (srov. Hoover 2003; Grieve 2007).

5 Viz Fleay 1874a, 1874b, 1874c, 1874d, 1876; Furnivall 1874b; 1874c.

Krom bohatosti slovníku navrhl ovšem Yule v *Statistical Study* ještě další metodu založenou na četnosti iniciálních grafémů slovních typů. Přestože v dnešních přehledových publikacích bývá připomínána jen výjimečně, můžeme ji považovat za průkopnickou a předjímající přístupy mnohem pozdější. Podrobněji se jí budeme věnovat v kap. 1.3.3.

1.3 Multidimenzionální analýzy

Za největší zlom ve stylometrii 20. století lze bezesporu označit studii Fredericka Mostellera a Davida L. Wallace (1964) věnovanou autorství tzv. *Listů federalistů* (*Federalist Papers*), která patří dodnes mezi nejcitovanější práce v oboru. Mosteller a Wallace vrací do hry dnes obecně přijímaný předpoklad formulovaný W. B. Smithem (srov. kap. 1.1), a sice že rysy využívané při rozpoznávání autorství musí být co možná nejméně závislé na tematice textu. Namísto do té doby využívaných autosémantik se proto rozhodli svou analýzu opřít naopak o nejfrekventovanější synsémantika a rozdíly v četnostech jejich variant (např. *while/whilst*). Především se ale jejich analýza na rozdíl od starších přístupů nezakládala na porovnání izolovaných četností ale na vzájemném srovnání celých sad sledovaných rysů. Tím začíná významný obrat od jednoduchých univariačních metod k složitějším multidimenzionálním analýzám, který vedl v 80. letech k uplatnění takových statistických metod jako vícerozměrná analýza rozptylu (Larsen–Rencher–Layton 1980) nebo analýza hlavních komponent (Burrows–Hassal 1987; Burrows 1989). Zdaleka největšího úspěchu pak – přinejmenším v oblasti rozpoznávání autorství beletristických textů – dosáhla tzv. Burrowsova míra Delta.

1.3.1 Burrowsova Delta

Tzv. míru Delta (Δ) navrhl John F. Burrows (2002, 2003) jako univerzální metriku stylistické podobnosti mezi texty primárně určenou pro situace, kdy existuje text, jehož autorství je neznámé nebo bylo zpochybněno (sporný text: t_0), a konečná množina potenciálních autorů a jimi produkované texty (kandidátská množina: $T = \{t_1, t_2, t_3, \ldots, t_m\}$). Za autora sporného textu je pak označen ten z kandidátů, jehož text vykazuje největší podobnost se sporným textem (tj. nejnižší hodnotu Delta).

Stejně jako Mosteller a Wallace vychází i Burrows z vysokofrekvenčního pásma lexika; výpočet Delty je založen na velikosti rozdílů mezi frekvencemi nejčetnějších slov:

(1) Z celého korpusu (tj. $t_0 \cup T$) je vybráno n nejfrekventovanějších slov $w_1, w_2, w_3, \ldots, w_n$.

(2) Každý text $t_a \in \{t_0, t_1, t_2, \ldots, t_m\}$ je reprezentován vektorem $\mathbf{f}_a = (f_1(t_a), f_2(t_a), \ldots, f_n(t_a))$, kde $f_i(t_a)$ značí relativní frekvenci slova w_i v t_a.

(3) Vzhledem k tomu že v každém textu psaném přirozeným jazykem jsou frekvence několika málo nejčetnějších slov násobně vyšší než frekvence slov z dalších ranků (1. Zipfův zákon), bývají i rozdíly mezi frekvencemi těchto nejčetnějších slov v různých textech řádově větší než u slov následujících. Pokud by tak byla Delta vypočtena přímo z rozdílu mezi relativními frekvencemi, nebyla by výsledná hodnota ovlivněna prakticky ničím jiným než pár slovy z předních míst rankové frekvenční distribuce. Relativní frekvence slov jsou proto napříč korpusem transformovány na z-skóry: $\mathbf{t}_a = (z_1(t_a), z_2(t_a), \ldots, z_n(t_a))$, kde:

$$z_i(t_a) = \frac{f_i(t_a) - \mu_i}{\sigma_i}$$

$$\mu_i = \frac{1}{m}\sum_{j=0}^{m} f_i(t_j) \tag{1.2}$$

$$\sigma_i = \sqrt{\frac{1}{m}\sum_{j=0}^{m}(f_i(t_j) - \mu_i)^2}$$

Tímto způsobem získáváme u frekvencí každého slova napříč korpusem distribuci s průměrem 0 a směrodatnou odchylkou 1.

(4) Stylistická odlišnost (Δ) mezi texty t_a a t_b je spočtena jako aritmetický průměr absolutních hodnot rozdílů z-skórů jednotlivých slov:

$$\Delta(t_a, t_b) = \frac{\sum_{i=1}^{n} \left| z_i(t_a) - z_i(t_b) \right|}{n} \tag{1.3}$$

(5) Za autora textu t_0 je označen ten z kandidátů, jehož text $t_a \in T$ vykazuje nejnižší hodnotu $\Delta(t_a, t_0)$.

Jako konkrétní ilustraci principu výpočtu Burrowsovy Delty (n = 50) uvádíme modelovou situaci, kdy sporný text představuje sbírka *Hudba v duši* (1886) Jaroslava Vrchlického a kandidátskou množinu *Zlomky Epopeje* (1886) téhož autora a *Nové písně* (1888) Svatopluka Čecha: OBR. 1.3 zobrazuje relativní četnosti 50 nejfrekventovanějších slov v těchto třech sbírkách, OBR. 1.4 zobrazuje tato data transformovaná na z-skóry, OBR. 1.5 zobrazuje absolutní hodnoty rozdílů mezi z-skóry ve sbírkách z kandidátské množiny a z-skóry ve sporném textu; poslední dva sloupce udávají jejich průměrnou hodnotu (Δ).

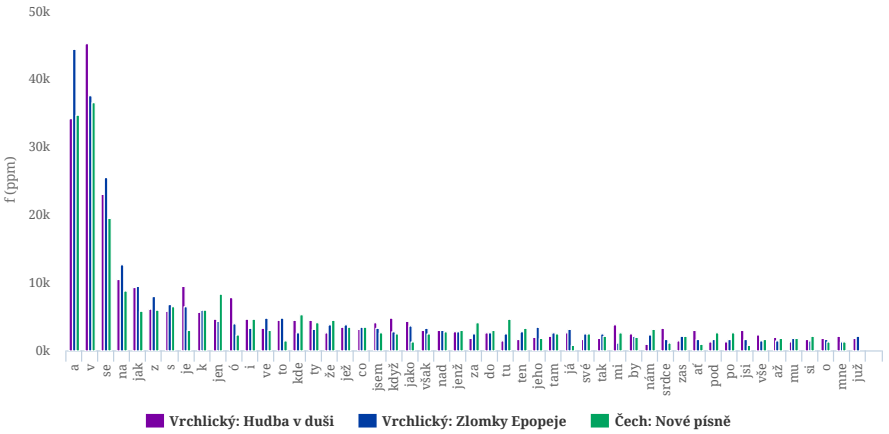
Díky vysoké úspěšnosti a zároveň jednoduchému a intuitivnímu principu se Delta rychle stala populárním a hojně užívaným nástrojem pro určování autorství. V následujících letech byla navržena řada drobných modifikací (např. Hoover 2004a, 2004b), podstatný krok vpřed ale přinesla interpretace principu fungování Delty Shlomo Argamona.

1.3.2 Geometrická interpretace a modifikace Burrowsovy Delty

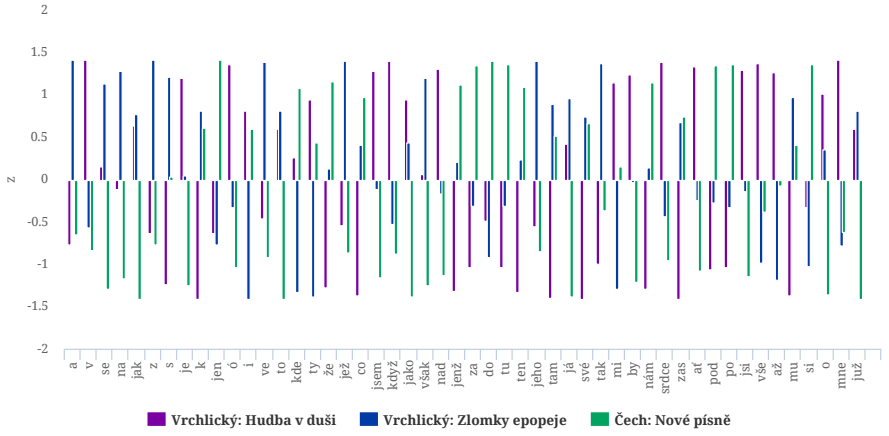
Argamon (2008) upozornil na to, že míra Delta, kterou Burrows navrhl zcela intuitivně, je ve skutečnosti ekvivalentem *manhattanské vzdálenosti* mezi dvěma vektory a celá metoda tak může být pokládána za klasifikaci metodou nejbližšího souseda, resp. za speciální případ populárního klasifikačního algoritmu k-nearest neighbor, kde k = 1.

Argamon vychází z jednoduché aritmetické úvahy: slouží-li Delta pouze k seřazení kandidátů, je dělení sumy rozdílů počtem analyzovaných slov (n) irelevantní, protože n je pro všechny texty konstantní a výsledné pořadí nijak neovlivní. Po jeho vypuštění z jmenovatele vzorce 1.3 dostáváme prostou sumaci absolutních hodnot rozdílů jednotlivých z-skórů, tj. manhattanskou vzdálenost (D_{M}) vektorů $\mathbf{t_a}$ a $\mathbf{t_b}$ (srov. OBR. 1.6, OBR. 1.7):

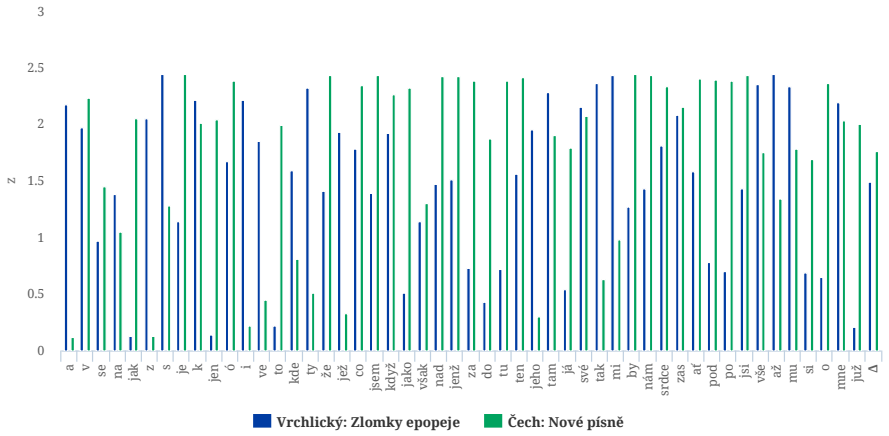
$$\Delta(t_a, t_b) \propto D_{\mathrm{M}}(\mathbf{t_a}, \mathbf{t_b}) = \sum_{i=1}^{n} \left| z_i(t_a) - z_i(t_b) \right| \tag{1.4}$$



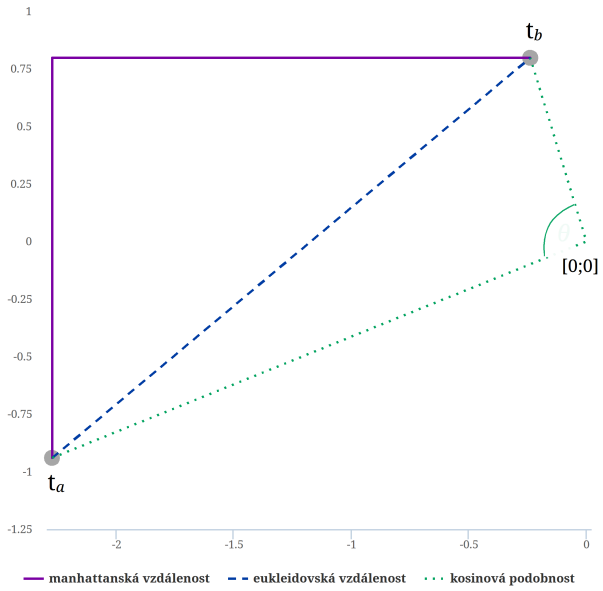
OBR. 1.3: Ilustrace výpočtu Burrowsovy Delty pro Vrchlického *Hudbu v duši* (sporný text), Vrchlického *Zlomky epopeje* a Čechovy *Nové písně* (kandidátská množina). Relativní frekvence 50 nejčetnějších slov.



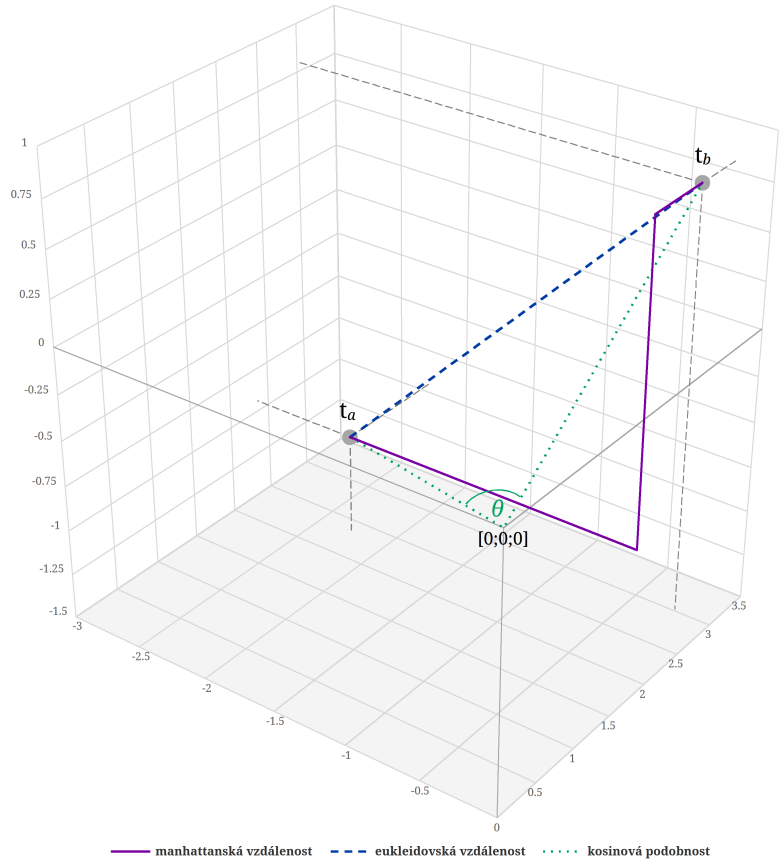
OBR. 1.4: Ilustrace výpočtu Burrowsovy Delty pro Vrchlického *Hudbu v duši* (sporný text), Vrchlického *Zlomky epopeje* a Čechovy *Nové písně* (kandidátská množina). Relativní frekvence 50 nejčetnějších slov transformované na *z*-skóry.



OBR. 1.5: Ilustrace výpočtu Burrowsovy Delty pro Vrchlického *Hudbu v duši* (sporný text), Vrchlického *Zlomky epopeje* a Čechovy *Nové písně* (kandidátská množina). Absolutní hodnoty rozdílů mezi *z*-skóry ve sbírkách z kandidátské množiny a *z*-skóry ve sporném textu; poslední dva sloupce udávají jejich průměrnou hodnotu.



OBR. 1.6: Ilustrace manhattanské vzdálenosti, eukleidovské vzdálenosti a kosinové podobnosti vektorů $\mathbf{t}_a$ a $\mathbf{t}_b$ v dvourozměrném prostoru.



OBR. 1.7: Ilustrace manhattanské vzdálenosti, eukleidovské vzdálenosti a kosinové podobnosti vektorů $\mathbf{t}_a$ a $\mathbf{t}_b$ v trojrozměrném prostoru.

V tomtéž článku navrhl Argamon i modifikaci Burrowsovy metody – tzv. kvadratickou Deltu (Δ_Q) – založenou na eukleidovské vzdálenosti (D_E) mezi danými vektory:

$$D_E(\mathbf{t}_a, \mathbf{t}_b) = \sqrt{\sum_{i=1}^{n} (z_i(t_a) - z_i(t_b))^2} \quad (1.5)$$

Vzhledem k tomu, že ani odmocnění celého výrazu neovlivňuje pořadí výsledných hodnot, definuje Argamon pro jednoduchost výpočtu Δ_Q jako čtverec eukleidovské vzdálenosti:

$$\Delta_Q(t_a, t_b) = \sum_{i=1}^{n} (z_i(t_a) - z_i(t_b))^2 \quad (1.6)$$

Druhou často užívanou modifikaci Burrowsovy metody (Smith–Aldridge 2011) představuje tzv. kosinová Delta ($\Delta_\angle$) vycházející z kosinové podobnosti vektorů, tedy z kosinu jimi svíraného úhlu θ:

$$\cos(\theta) = \frac{\sum_{i=1}^{n} z_i(t_a) z_i(t_b)}{\sqrt{\sum_{i=1}^{n} z_i(t_a)^2} \sqrt{\sum_{i=1}^{n} z_i(t_b)^2}} \quad (1.7)$$

Protože kosinus úhlu může nabývat hodnoty mezi 1 (stejný směr vektorů, maximální podobnost) a −1 (opačný směr vektorů, maximální nepodobnost) je vhodnější výpočet upravit tak, aby (stejně jako u Δ a Δ_Q) odpovídala nižší hodnota větší podobnosti a naopak, proto:

$$\Delta_\angle(t_a, t_b) = 1 - \cos(\theta) \quad (1.8)$$

Výše popsané metriky z rodiny Delta byly testovány napříč jazyky i typy textů, a to jak s různým nastavením počtu analyzovaných jednotek (n), tak s řadou jiných analyzovaných rysů jako frekvence lemmat, znakových n-gramů nebo slovních n-gramů (viz např. Eder 2011; Rybicki–Eder 2011; Jannidis et al. 2015). Empiricky při tom bylo zjištěno, že:

(1) Úspěšnost Delty závisí v první řadě na velikosti analyzovaných vzorků. Na základě testování Delty na několika jazycích odhadnul Eder (2015) minimální velikost vzorků na 5000 tokenů. Toto stanovisko ale později (2017) revidoval s tím, že minimální velikost závisí na řadě faktorů a nelze ji proto stanovit obecně („telling apart Hemingway and Dickens will always be easier than distinguishing Bronte sisters“, Eder 2017: 1).

(2) Úspěšnost Delty bývá vyšší u prózy než u básnických textů (Rybicki–Eder 2011).

(3) Nejfrekventovanější slova obvykle vykazují ve flektivních jazycích nižší atribuční sílu než nejfrekventovanější znakové n-gramy. Kestemont (2014) podává smysluplné vysvětlení: zatímco v angličtině a jiných převážně izolačních jazycích nesou nejfrekventovanější slova převážně gramatický význam (tedy kontextově nezávislou informaci), ve flektivních jazycích je gramatická informace z velké části „skryta“ v gramatických morfémech, které většinou nebývají realizovány jako samostatná slova. Znakové n-gramy mohou ale při vhodném nastavení n ve flektivních jazycích tuto informaci poměrně dobře aproximovat (např. při $n = 4$ a zachování mezer („_“) budou v češtině jako samostatné jednotky extrahovány jak dvoupísmenná synsémantika („_na_“), tak afixy typu „_roz“, „ých_“, „ého_“).

1.3.3 Yuleho metoda iniciálních grafémů

Ještě než se podíváme na další moderní techniky určování autorství, vraťme se k výše zmíněné metodě G. U. Yuleho (kap. 1.2) založené na četnosti iniciálních grafémů. Ačkoliv z dnešního pohledu se jedná o metodu zastaralou a nepříliš spolehlivou, z hlediska způsobu zpracování dat si pozornost zaslouží.

Sám Yule (1944: 183) uvádí, že k této metodě dospěl vlastně náhodou, když pro potřeby experimentu založeného na bohatosti slovníku zaznamenával četnosti slov. Pro dílo každého analyzovaného autora vytvořil abecedně řazenou kartotéku, v níž každému slovu odpovídala karta s údajem o jeho frekvenci. Když jednou zároveň otevřel zásuvky obsahující údaje o díle Johna Bunyana a Thomase Macaulaye, všiml si, že přestože obě sady obsahovaly zhruba stejný počet karet, umístění abecedních oddělovačů se značně lišilo. To přivedlo Yulea k myšlence, že by četnosti iniciálních grafémů jednotlivých lemmat (počet karet mezi jednotlivými oddělovači) bylo možné využít jako vhodný rys pro rozpoznání autorství.

Testování atribuční síly tohoto ukazatele provedl Yule na vzorcích extrahovaných z Bunyanova a Macaulayho korpusu. Sledoval při tom, jestli frekvence iniciálních grafémů převedené na pořadí (ranky) v těchto vzorcích zachovávají větší podobnost s hodnotami zjištěnými v korpusu, z něhož pocházejí, než s hodnotami zjištěnými v korpusu druhého autora. Podstatné ale je, že se Yule nespokojil s (tehdy běžným) porovnáváním izolovaných hodnot ale použil pro tyto potřeby de facto *klasifikaci metodou nejbližšího souseda*. Yule popisuje metodu následovně:

> „We write down the differences of the ranks in Bunyan sample A from the ranks in the total Bunyan vocabulary, paying no attention to sign; the sum at the foot is a rough measure of the badness of agreement between the sample ranking and for the total of Bunyan vocabulary. In exactly same way we enter […] the differences between the sample A ranking and the ranking for the total Macaulay vocabulary, and enter the sum, without regard to sign, at the foot. These respective sums are 10 and 37: we have found that the ranking of the given sample differs much less from that of the Bunyan vocabulary than from that of the Macaulay vocabulary, and are left in practically no doubt that the given sample (if we did not know from which author it had come) should be assigned to Bunyan“ (Yule 1944: 190).

K tomu ještě připojuje sofistikovanější způsob založený na Spearmanově koeficientu pořadové korelace (ibid.: 191).

Lze ukázat, že první postup je vlastně klasifikací metodou nejbližšího souseda na základě *manhattanské metriky*, druhý postup pak klasifikací metodou nejbližšího souseda na základě *eukleidovské metriky*: máme vzorek s, Bunyanův korpus c_B, Macaulayův korpus c_M a 26 grafémů anglické abecedy g_1, g_2, …, g_{26}. Vzorek s a oba korpusy jsou reprezentovány vektory

$$\mathbf{s} = (r_1(s), r_2(s), \ldots, r_{26}(s)),$$

$$\mathbf{c_B} = (r_1(c_B), r_2(c_B), \ldots, r_{26}(c_B)),$$

$$\mathbf{c_M} = (r_1(c_M), r_2(c_M), \ldots, r_{26}(c_M)),$$

kde $r_i(x)$ značí pořadí iniciálního grafému g_i ve frekvenční rankové distribuci vzorku/korpusu x.

První postup pak nedělá nic jiného, než že přiřazuje vzorku *s* toho autora *a*, jehož korpus je z hlediska manhattanské metriky (D_{M}) vzorku bližší.

$$D_{\mathrm{M}}(\mathbf{s},\mathbf{c_a}) = \sum_{i=1}^{n}\left|r_i(s)-r_i(c_a)\right| \quad (1.9)$$

U druhého postupu lze ukázat, že klasifikace na základě Spearmanova koeficientu pořadové korelace (vzorec 1.10) je totožná s klasifikací metodou nejbližšího souseda za použití eukleidovské metriky (vzorec 1.11).

$$\rho(\mathbf{s},\mathbf{c_a}) = 1 - \frac{6\sum_{i=1}^{n}(r_i(s) - r_i(c_a))^2}{n(n^2-1)} \quad (1.10)$$

$$D_{\mathrm{E}}(\mathbf{s},\mathbf{c_a}) = \sqrt{\sum_{i=1}^{n}(r_i(s) - r_i(c_a))^2} \quad (1.11)$$

Obecně tedy platí:

$$\rho = -D_{\mathrm{E}}^2\frac{6}{n(n^2-1)} + 1 \quad (1.12)$$

Vzhledem k tomu, že *n* je konstanta (počet dimenzí vektorového prostoru, v tomto případě 26 grafémů anglické abecedy), je pořadí kandidátů autorství určené stoupající hodnotou Spearmanova koeficientu totožné s pořadím kandidátů určeným klesající hodnotou eukleidovské vzdálenosti.

Metoda nejbližšího souseda se tedy u Yuleho objevuje plných 60 let před tím, než byla Burrowsem zpopularizována ve stylometrii, a 10–20 let před tím, než se vůbec začala používat v datové analýze (srov. Pelillo 2014: 34).

1.4 Support Vector Machine

Vedle metrik z rodiny Delta se v posledních letech čím dál tím víc prosazují i sofistikovanější metody strojového učení jako *random forest* (např. Tabata 2012), *naïve Bayes classifier* (např. Zhao–Zobel 2005) a zejména v dnešní době nejčastěji užívaná technika *support vector machine* (SVM; např. Diederich et al. 2003; Koppel–Schler 2004), na níž se v následující kapitole zaměříme.[6]

SVM je příkladem tzv. *eager learning*, což znamená, že algoritmus nejprve z označkovaných trénovacích dat odvozuje klasifikační funkci, kterou poté aplikuje při klasifikaci nových, dosud neznámých dat. Základní principy SVM lze ukázat na následujícím (uměle vytvořeném) příkladu: máme sporný text t_0 a 20 vzorků textů od obou kandidátů autorství (autor 1, autor 2). Všechny texty jsou reprezentovány *z*-skóry dvou nejfrekventovanějších slov („a“, „se“).

Během první fáze (učení) hledá SVM ve vektorovém prostoru nadrovinu (v našem případně dvourozměrných dat tedy přímku), která správně rozdělí data na vzorky autora 1

6 Následující výklad principů SVM vychází zejm. z Abney 2007; Burges 1998; Hearst 1998 a v neposlední řadě z názorných prezentací Alexandera Ihlera (University of California, Irvine): https://www.youtube.com/user/atihler

a vzorky autora 2. Během druhé fáze (klasifikace) je pak nadrovina využita ke klasifikaci sporného textu.

První graf na OBR. 1.8 ukazuje, že pokud lze takovou nadrovinu sestrojit, pak jich lze sestrojit nekonečně mnoho. V uvedeném příkladu by potom některé z nadrovin klasifikovaly sporný text jako dílo autora 1, jiné jako dílo autora 2. Ze všech těchto nadrovin vybírá SVM tu, která má maximální možnou vzdálenost k nejbližším vektorům z obou tříd (podpůrné vektory, viz druhý graf na OBR. 1.8). V tomto případě by byl tedy sporný text klasifikován jako text autora 1.

Obecně pro *n*-rozměrná data lze úlohu formulovat následovně: Máme trénovací data sestávající ze dvojic $(\mathbf{x}^{(1)}, y^{(1)})$, $(\mathbf{x}^{(2)}, y^{(2)})$, …, $(\mathbf{x}^{(m)}, y^{(m)})$, kde první člen představuje *n*-rozměrný vektor $\mathbf{x}^{(i)}=(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$ a druhý člen informaci o tom, ke které ze dvou tříd vektor náleží: $y^{(i)} \in \{-1, 1\}$. Hledáme normálový vektor $\mathbf{w}$ a parametr b určující nadrovinu

$$H: \ \mathbf{w}\cdot\mathbf{x} + b = 0 \qquad (1.13),$$

která rozdělí vektorový prostor na dva podprostory tak, že každý podprostor bude obsahovat pouze data z jedné třídy a vzdálenost k nejbližšímu vektoru bude maximální možná.

Tyto požadavky lze formálně definovat prostřednictvím orientované vzdálenosti d vektorů $\mathbf{x}^{(i)}$ od nadroviny H, která nabývá kladné hodnoty u vektorů v jednom podprostoru nadroviny H, záporné hodnoty u vektorů v druhém podprostoru:

$$d(\mathbf{x}^{(i)}, H) = \frac{\mathbf{x}^{(i)}\cdot\mathbf{w}+b}{\|\mathbf{w}\|} \qquad (1.14)$$

Máme-li dvě třídy $y^{(i)} \in \{-1, 1\}$, můžeme tak požadavek, aby každý podprostor obsahoval pouze vektory jedné třídy formulovat jako

$$\begin{aligned} \forall i: y^{(i)}=1, \qquad & \frac{\mathbf{x}^{(i)}\cdot\mathbf{w} + b}{\|\mathbf{w}\|} > 0 \\ \forall i: y^{(i)}=-1, \qquad & \frac{\mathbf{x}^{(i)}\cdot\mathbf{w} + b}{\|\mathbf{w}\|} < 0 \end{aligned} \qquad (1.15),$$

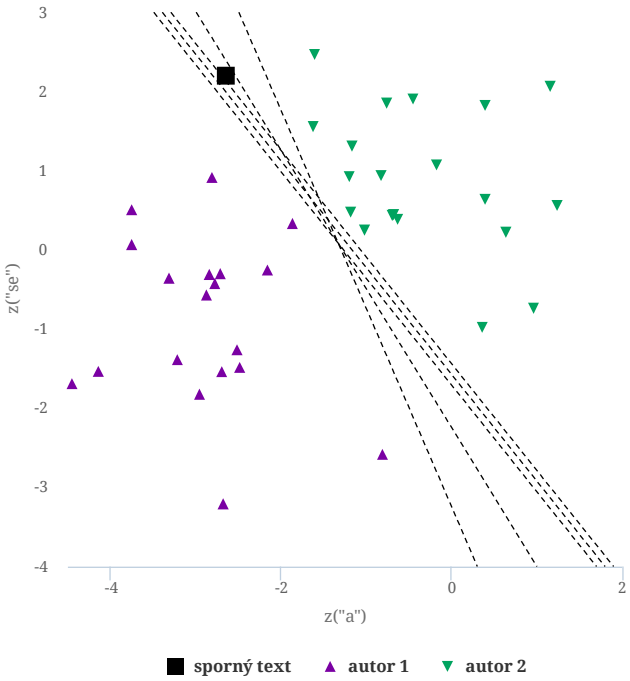
což lze zjednodušit na

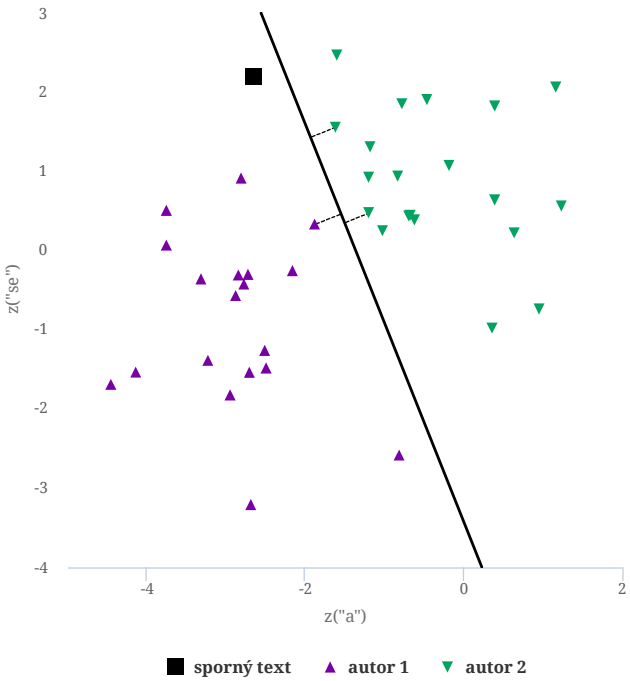
$$\forall i, \qquad y^{(i)}\frac{\mathbf{x}^{(i)}\cdot\mathbf{w} + b}{\|\mathbf{w}\|} > 0 \qquad (1.16)$$

Dále požadujeme, aby šířka hraničního pásma byla co největší možná; chceme tedy maximalizovat eukleidovskou (neorientovanou) vzdálenost nejbližších (podpůrných) vektorů od nadroviny H. Celou úlohu bychom tak mohli formulovat jako

$$\begin{aligned} & \max_{\mathbf{w},b} \min_{i} \left| \frac{\mathbf{x}^{(i)}\cdot\mathbf{w} + b}{\|\mathbf{w}\|} \right| \\ & \text{kde} \ \forall i, \qquad y^{(i)}\frac{\mathbf{x}^{(i)}\cdot\mathbf{w} + b}{\|\mathbf{w}\|} > 0 \end{aligned} \qquad (1.17)$$

Taková úloha má ale nekonečně mnoho řešení, protože požadavky specifikují pouze směr vektoru $\mathbf{w}$ a nikoliv jeho velikost $\|\mathbf{w}\|$. Z praktických důvodů se proto požaduje, aby





OBR. 1.8: Ilustrace principu SVM (umělá data). Nahoře: různé nadroviny oddělující trénovací data autora 1 a autora 2. Dole: Nadrovina s nejširším možným hraničním pásmem; přerušované čáry ukazují vzdálenosti k podpůrným vektorům.

velikost $\|\mathbf{w}\|$ byla nepřímo úměrná eukleidovské vzdálenosti podpůrných vektorů od nadroviny *H*, tj.

$$\frac{1}{\|w\|} = \min_i \left| \frac{\mathbf{x}^{(i)} \cdot \mathbf{w} + b}{\|\mathbf{w}\|} \right| \qquad (1.18)$$

Pro podpůrné vektory pak totiž můžeme zjednodušit požadavek na

$$\left| \mathbf{x}^{(i)} \cdot \mathbf{w} + b \right| = 1 \qquad (1.19)$$

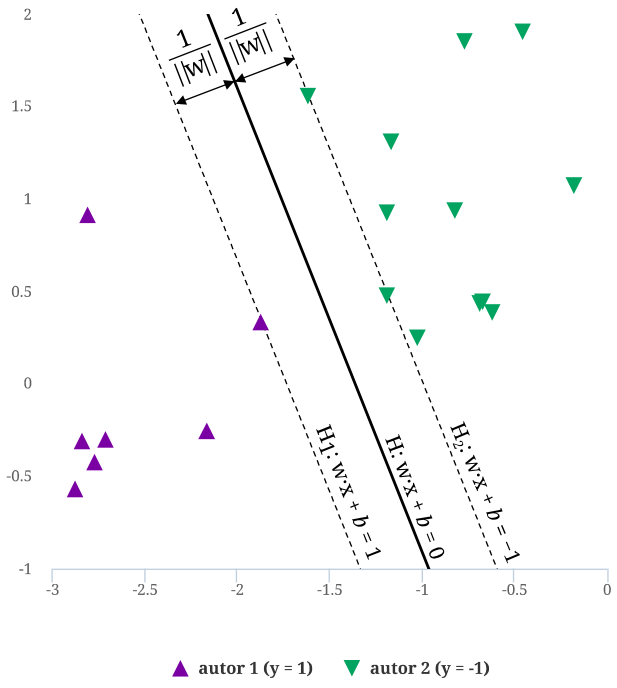
Pro všechny vektory na

$$\forall i, \qquad y^{(i)}(\mathbf{x}^{(i)} \cdot \mathbf{w} + b) \geq 1 \qquad (1.20)$$

Tím se dostáváme k základní formulaci optimalizačního problému SVM: hledáme-li normálový vektor **w** a parametr *b* určující nadrovinu *H*, která má nejširší možné hraniční pásmo, a má-li být šířka hraničního pásma nepřímo úměrná velikosti **w** (vzorec 1.18), je řešením nejmenší možný normálových vektor **w**, pro nějž platí nerovnice 1.20 (viz OBR. 1.9). Z praktických důvodů se místo velikosti $\|\mathbf{w}\|$ minimalizuje polovina její druhé mocniny, tedy:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{kde } \forall i, \qquad y^{(i)}(\mathbf{x}^{(i)} \cdot \mathbf{w} + b) \geq 1 \qquad (1.21)$$

Tato úloha je pak řešena pomocí Lagrangeových multiplikátorů (postup viz např. Abney 2007: 117–119).



OBR. 1.9: Ilustrace principu SVM.

Tento příklad představuje nejjednodušší možnou variantu klasifikace *n*-rozměrných dat. V praxi je ovšem často třeba pracovat jednak s daty, která nejsou lineárně separovatelná, jednak s klasifikací do více než dvou tříd.

1.4.1 Lineárně neseparovatelná data

U dat, pro která neexistuje žádná nadrovina, která by dané třídy separovala, existují dva základní způsoby řešení: (1) zmírnění podmínky nepropustnosti hraničního pásma a (2) jádrová transformace původních dat do vyšších dimenzí.

(1) *SVM s propustným hraničním pásmem* (soft margin SVM), který je využíván především u relativně málo zašuměných dat, ruší podmínku požadující, aby každý podprostor obsahoval pouze vektory jedné třídy, a místo ní zavádí přídatnou proměnnou ξ, která penalizuje vektory nacházející se „na špatné straně“ nadroviny. Cílem tedy je nalézt takovou nadrovinu, která má nejširší možné hraniční pásmo a zároveň minimalizuje „přesahy“ vektorů jedné třídy do podprostoru třídy druhé.

Pro vektory $\mathbf{x}^{(i)}$ nacházející se v podprostoru druhé třídy je ξ_i definováno jako eukleidovská vzdálenost $\mathbf{x}^{(i)}$ od toho okraje hraničního pásma, na němž leží podpůrné vektory jeho třídy ($H_{y^{(i)}}$), normalizovaná šířkou hraničního pásma (viz OBR. 1.10). Pro tyto vektory tak platí:

$$\xi_i = \frac{\left|\frac{\mathbf{w}\cdot x^{(i)}+b-y^{(i)}}{\|\mathbf{w}\|}\right|}{\frac{1}{\|\mathbf{w}\|}} = \left|\mathbf{w}\cdot x^{(i)}+b-y^{(i)}\right| \tag{1.22}$$

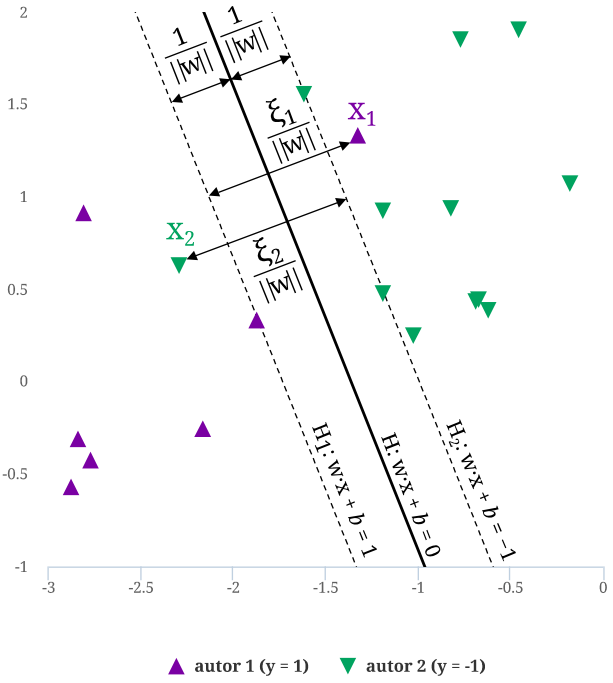
Pro ostatní vektory $\xi_i = 0$.

Optimalizační problém SVM (vzorec 1.21) je pak rozšířen do podoby

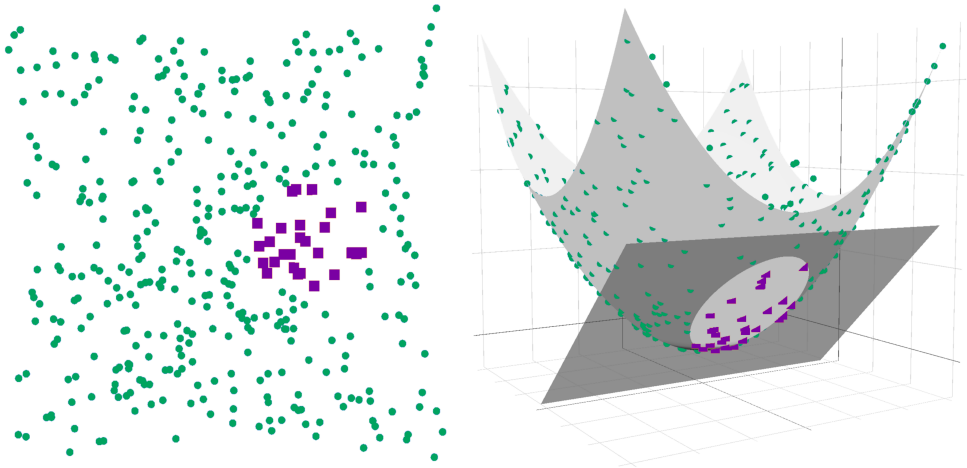
$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \quad \text{kde } \forall i, \quad y^{(i)}(\mathbf{x}^{(i)}\cdot\mathbf{w} + b) \geq 1 - \xi_i \quad \wedge \quad \xi_i \geq 0 \tag{1.23},$$

kde C představuje volitelný parametr modelu (váha, kterou přikládáme penalizaci chyb).

(2) U více zašuměných lineárně neseparovatelných dat bývá využívána transformace původně *n*-rozměrných dat do (*n*+*k*)-rozměrného prostoru, kde jsou již data lineárně separovatelná (*kernel trick*). Příkladem je transformace původně dvourozměrných dat (OBR. 1.11 vlevo) do trojrozměrného prostoru (OBR. 1.11 vpravo), kde každému původnímu vektoru $\mathbf{x}=(x_1, x_2)$ odpovídá vektor $\mathbf{x'}=(x_1, x_2, x_1^2+x_2^2)$. U lingvistických dat, kde obvykle na relativně malý počet tříd připadá velmi vysoký počet dimenzí, ale není většinou třeba kernel trick aplikovat.



OBR. 1.10: Ilustrace SVM s propustným hraničním pásmem.



OBR. 1.11: Ilustrace jádrové transformace lineárně neseparovatelných dvourozměrných dat. Transformační funkce: $\Phi(a, b) = (a, b, a^2 + b^2)$.

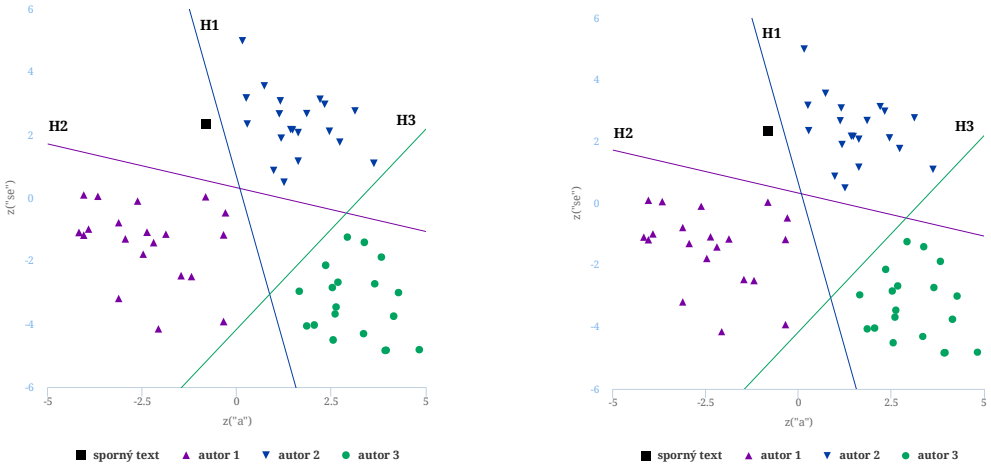
1.4.2 Klasifikace do více tříd

Protože SVM je (jak jsme viděli) z principu binární klasifikátor, je tato situace řešena rozdělením problému na více binárních úloh, a to buď přístupem *one-vs.-rest* nebo *one-vs.-one*.

(1) V případě *one-vs.-rest* je pro každou třídu vytvořena klasifikační funkce, která ji separuje od ostatních dat (pro data o k třídách tedy k klasifikačních funkcí). Pokud ze všech k klasifikačních funkcí pouze jedna přiřazuje sporný text konkrétnímu autorovi a všechny ostatní ho přiřazují ke sloučeným „zbytkovým datům“, je spornému textu přiřazen tento autor. Existuje-li více klasifikačních

funkcí, které přiřazují sporný text konkrétnímu autorovi, rozhoduje větší vzdálenost od příslušné nadroviny. V příkladu na OBR. 1.12 vlevo by tak byl sporný text přiřazen autorovi 2.

(2) V případě *one-vs.-one* je klasifikační funkce vytvořena pro každou dvojici tříd (pro data o k třídách tedy $\frac{k(k-1)}{2}$ klasifikačních funkcí). Každá z těchto funkcí přiřazuje sporný text jednomu autorovi. Ve výsledku je sporný text přiřazen tomu autorovi, pro něhož hlasoval největší počet dílčích klasifikátorů.



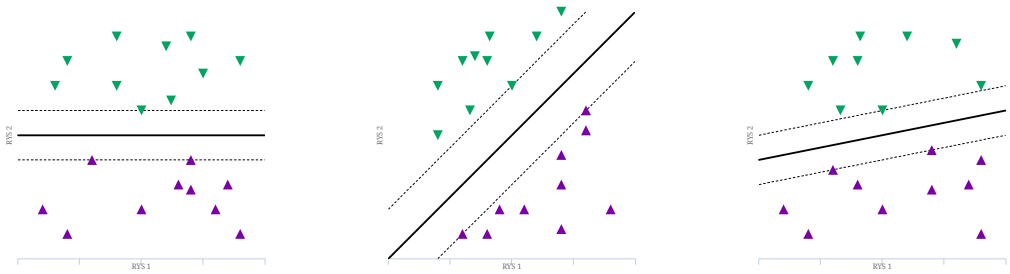
OBR. 1.12: Ilustrace klasifikace do více tříd pomocí SVM. Vlevo: přístup one-vs.-rest; vpravo: přístup one-vs.-one. V obou případech je sporný text přiřazen autorovi 2: v případě one-vs.-rest sporný text sice nadrovina H_1 přiřazuje i autorovi 1, vzdálenost od nadroviny H_2 je ale větší; v případě one-vs.-one sporný text autorovi 2 přiřazují dvě nadroviny ($H_{1,2}$; $H_{2,3}$), autorovi 1 pouze jedna ($H_{1,3}$).

1.4.3 Normálový vektor nadroviny jako ukazatel klasifikační síly rysů

Užitečnou vlastností nadroviny konstruované metodou SVM je možnost usuzovat ze souřadnic jejího normálového vektoru na míru, s jakou jednotlivé rysy přispívají ke správné klasifikaci.

Zůstaňme pro jednoduchost u dvourozměrných dat (rys 1, rys 2): máme-li nadrovinu dvourozměrného prostoru (tj. přímku) určenou obecnou rovnicí $w_1x_1 + w_2x_2 + b = 0$, pak normálový vektor $\mathbf{w} = (w_1, w_2)$ určuje, jak známo, její sklon, zatímco parametr b její vertikální posunutí. A právě sklon nadroviny lze interpretovat jako ukazatel klasifikační síly.

Vyjděme z krajních případů, kdy jeden z rysů nepřináší vůbec žádnou informaci: pokud je informační hodnota rysu 1 (osa x_1) nulová a celá klasifikace je provedena pouze na základě hodnot rysu 2 (osa x_2), je dělící nadrovina s osou x_1 rovnoběžná, tzn. $w_1 = 0$ (OBR. 1.13 vlevo). (V opačném případě, kdy by nulovou informační hodnotu vykazoval rys 2, byla by výsledkem nadrovina rovnoběžná s osou x_2, tzn. $w_2 = 0$.) Situaci, kdy oba rysy přispívají ke klasifikaci stejnou měrou, odpovídá nadrovina určená rovnicí $x_2 = x_1 + b$, tzn. $w_1 = w_2$ (OBR. 1.13 uprostřed). Jakýkoliv jiný sklon nadroviny pak lze interpretovat jako vyšší klasifikační sílu jednoho z rysů: např. OBR. 1.13 vpravo znázorňuje situaci, kterou lze intuitivně chápat tak, že rys 1 přispívá ke správné klasifikaci méně než rys 2 ($w_1 < w_2$).



OBR. 1.13: Ilustrace klasifikační síly rysů. Směrový vektor nadroviny: $\mathbf{w} = (w_1, w_2)$. Vlevo: klasifikační síla rysu 2 je nulová, klasifikace probíhá pouze na základě rysu 1 ($w_1 = 0$); uprostřed: klasifikační síla obou rysů je stejná ($w_1 = w_2$); vpravo: klasifikační síla rysu 2 je větší než klasifikační síla rysu 1 ($w_1 < w_2$) .

Stejným způsobem lze interpretovat směrové vektory nadrovin i u dat více než dvourozměrných. Dodejme ale, že směrový vektor lze tímto způsobem využít pouze u lineárních variant SVM. Po transformaci n-rozměrných dat do $(n+k)$-rozměrného prostoru (kap. 1.4.1.(2)) není vztah mezi normálovým vektorem a jednotlivými rysy smysluplně interpretovatelný.

1.4.4 Validace

Základní metodu testování úspěšnosti modelu vytvořeného strojovým učením obecně představuje tzv. *hold-out validace*, kdy jsou data (obvykle náhodně) rozdělena na trénovací a testovací množinu (obvykle v poměru 2 : 1). Na první z nich je model natrénován, na datech z druhé je provedena klasifikace a zaznamenáno, jak velká část vzorků byla klasifikována správně. Tento údaj pak slouží jako odhad úspěšnosti modelu.

Přesnější odhad lze získat *křížovou validací*, kdy jsou data rozdělena do k stejně velkých částí. Jedna část je potom vyňata, ostatních $k - 1$ částí využito coby trénovací množina, a následně je na vyňaté části provedena klasifikace. Tento proces je opakován pro každou z částí, čehož výsledkem je k odhadů úspěšnosti. Výsledný odhad úspěšnosti je pak dán jako aritmetický průměr výstupů z jednotlivých iterací.

U dat, kde jednotlivé třídy obsahují relativně málo vzorků (což je obvykle případ lingvistických dat) bývá upřednostňována *leave-one-out křížová validace*, kdy jsou data sestávající z n vzorků rozdělena na $k = n$ částí – v každé iteraci je tedy model testován na jediném vzorku a výsledný odhad úspěšnosti je dán četností správných klasifikací.

Samotný výsledek validace má ovšem jen malou výpovědní hodnotu. Abychom mohli klasifikátor považovat za funkční, je především třeba, aby jeho úspěšnost překročila určitý práh (*baseline*), kterého lze dosáhnout prostým hádáním. Dosáhne-li například binární klasifikátor 90% úspěšnosti na datech, kde 90 % případů náleží jedné třídě, není jeho využitelnost příliš vysoká vzhledem k tomu, že stejné úspěšnosti lze dosáhnout triviálním klasifikátorem, který by každému vzorku přiřazoval nejfrekventovanější třídu. Jednu ze základních metod určení takového prahu (vedle právě zmíněné, která je ovšem vhodná především pro nevyvážené datasety) představuje *random baseline* (RB), tj. nejpravděpodobnější úspěšnost takového klasifikátoru, který přiřazuje vzorkům autory zcela náhodně:

$$\mathrm{RB} = \sum_{i=1}^{N} \left(\frac{n_a}{X}\right)^2 \qquad (1.24)$$

kde N značí počet tříd, X značí počet vzorků a n_a značí počet vzorků v třídě a.

***

Z výše uvedeného můžeme odvodit hlavní výhody a nevýhody SVM oproti Burrowsově míře Delta a jejím modifikacím:

- Zatímco SVM přiděluje jednotlivým rysům různou váhu danou souřadnicemi normálového vektoru nadroviny (srov. kap. 1.4.3), u měr z rodiny Delta přispívá každý rys ke klasifikaci stejnou měrou. SVM by tedy měl být teoreticky více odolný vůči šumu. Příkladem může být OBR. 1.13 vlevo výše, kde SVM správně rozpoznává rys 1 jako pro klasifikaci irelevantní, zatímco míry z rodiny Delta by přikládaly oběma rysům stejnou váhu (Δ a $\Delta_{\angle Q}$ by tak chybně klasifikovaly levý podpůrný vektor u třídy označené fialovou barvou; jeho nejbližším sousedem je podpůrný vektor druhé třídy).
- Na druhou stranu SVM vyžaduje pro trénovací fázi relativně velký počet vzorků. Pro situace, kdy je v kandidátském korpusu k dispozici malý počet vzorků pro některé nebo všechny autory, jsou méně robustní míry z rodiny Delta teoreticky vhodnější.

1.5 Atribuce na základě versologických rysů

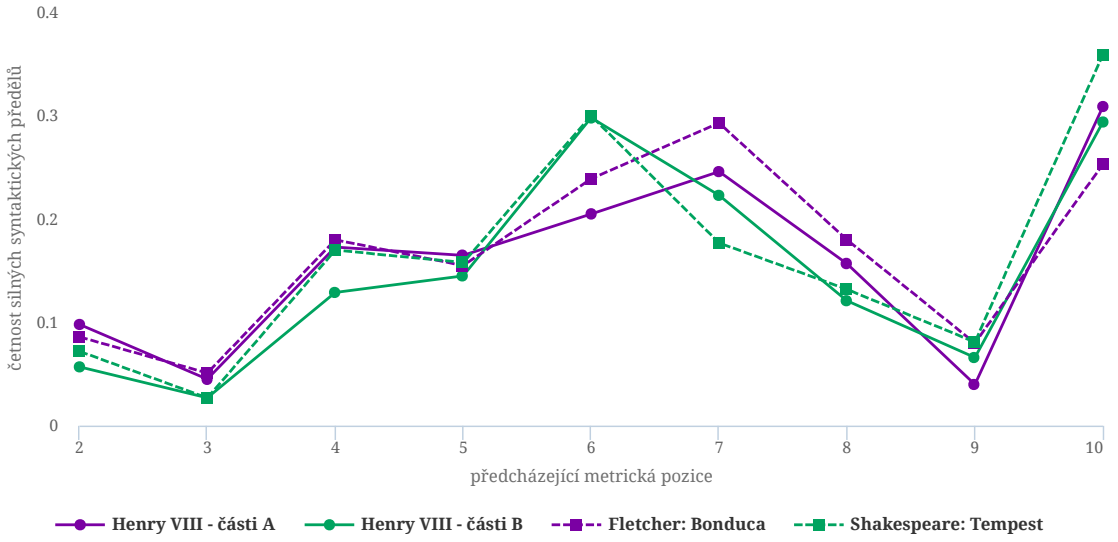
V předchozích kapitolách jsme viděli, že stylometrie využívala a využívá pro rozpoznání autorství jednak pestrou škálu technik, jednak pestrou škálu textových rysů. S versologickými rysy se ale – až na shakespearologický výzkum (viz kap. 1.1 a kap. 4.1.1) – nesetkáme. Výjimky ve 20. století představují atribuční pokusy samotných versologů, a to zejména versologů spojených s tzv. Ruskou školou.

Známým příkladem je studie Borise Tomaševského (1923/2008) zabývající se Puškinovou poémou *Rusalka*. Tato báseň byla dlouho považována za nedokončenou, v roce 1889 ale básník Dmitrij Zujev publikoval jím údajně nalezenou závěrečnou část. Tomaševskij změřil v „nalezeném“ textu četnosti přízvuků na jednotlivých metrických pozicích a četnosti mezislovních předělů mezi nimi. Na základě srovnání těchto měření s jinými prokazatelně Puškinovými texty pak nález prohlásil za podvrh.

Z dalších případů využití charakteristik verše a rýmu uveďme zpochybnění pravosti fragmentu 10 kapitoly *Evžena Oněgina* (Lotman–Lotman 1986), zpochybnění pravosti básní nově přidaných Alexandrem Iljušinem do souborného díla Gavrijila Batěnkova (Šapir 1997, 1998) a především četné práce Mariny Tarlinské o Shakespearovi a jeho současnících (zejm. Tarlinskaja 1987, 2014).

Tato versologická linie se ovšem vyvíjela zcela nezávisle na hlavním proudu stylometrie, za nímž metodologicky znatelně zaostala; zatímco stylometrie ušla – jak jsme viděli v předchozích kapitolách – během 20. století dlouhou cestu od jednoduchých textových charakteristik k sofistikovaným multidimenzionálním modelům, versologická linie zůstala u jednoduchých metod deskriptivní statistiky.

Můžeme to ilustrovat na příkladu z relativně nedávné knihy zmíněné M. Tarlinské *Shakespeare and the Versification of English Drama, 1561–1642* (2014: 140–149). Tarlinskaja se zde věnuje mimo jiné hře *Henry VIII.*, u níž od 19. století existuje hypotéza, že obsahuje části napsané Williamem Shakespearem (části A) a části napsané Johnem Fletcherem (části B).[7] Tuto hypotézu se snaží podpořit versologickými argumenty: poukazuje na to, že verše z části A se znatelně liší od veršů z částí B v distribuci „silných syntaktických předělů“ („strong syntactic breaks“).[8] Tarlinskaja změřila četnost těchto předělů ve verších částí A a částí B hry *Henry VIII.* a ve verších dvou dalších textů pocházejících z téhož období: Fletcherovy hry *Bonduca* a Shakespearovy hry *Tempest.* Zjištění že „silné syntaktické předěly“, odhlédneme-li od předělů po desáté slabice, se v částech A stejně jako v *Bonduce* objevují nejčastěji po sedmé slabice verše, zatímco v částech B a ve hře *Tempest* nejčastěji po šesté slabice verše (viz OBR. 1.14), pak pokládá za potvrzení platnosti výše zmíněné hypotézy.



OBR. 1.14: Četnosti „silných syntaktických předělů“ po jednotlivých slabikách (metrických pozicích) veršů částí A a částí B hry *Henry VIII.*, Fletcherovy hry *Bonduca* a Shakespearovy hry *Tempest*. Zdroj: Tarlinskaja 2014: tabulka B.3.

Stejným způsobem srovnává v částech hry *Henry VIII* a hrách *Tempest* a *Bonduca* Tarlinskaja četnosti monosylab na koncích mužských veršů, četnosti monosylab na koncích ženských veršů a četnosti tzv. enjambement (absence „silného syntaktického předělu“ na samotném konci verše). U všech těchto rysů konstatuje nápadnou podobnost mezi částmi A a *Bondukou* na jedné straně a částmi B a *Tempest* na straně druhé.

Jakkoliv se jedná o validní a relevantní argumenty, je třeba dodat, že co do způsobu zpracování je Tarlinské přístup v zásadě shodný s Mendenhallovým (srov. kap. 1.1) a že od té doby byly vyvinuty mnohem spolehlivější a robustnější metody než prosté konstatování podobnosti dvou výsledků měření.

7 Viz Speddingova a Ingramova atribuce (kap. 1.1) a celá kapitola věnovaná autorství hry (4.1).

8 „A strong syntactic break occurs, for example, at the juncture of sentences, or a sentence and a clause, […] between the author‘s and direct speech, […] or between a direct address and the rest of the utterance“ (Tarlinskaja 2014: 24).

1.6 Stylometrie v českém prostředí

Stejně jako u celé oblasti rozpoznávání autorství byly i počátky její kvantitativní větve v českém prostředí spojeny s otázkou autorství Rukopisů královédvorského a zelenohorského (RK, RZ).[9] V roce 1886, kdy kulminuje spor mezi zastánci a odpůrci pravosti, publikoval fyzik a astronom August Seydler v Masarykově *Athenaeu* pozoruhodný, dnes ovšem zřídka připomínaný článek *Počet pravděpodobnosti v přítomném sporu* (Seydler 1886). Seydler vychází z Gebauerova (1886) výčtu slovních tvarů objevujících se v RK, které nebyly doloženy v žádném středověkém textu, a naopak výčtu jejich výskytů v textech raného obrození („koincidence"), zejm. u pravděpodobného autora RK Václava Hanky. K těmto dokladům přistupuje de facto z perspektivy inferenční statistiky: vypočítává, že pravděpodobnost výskytu daných slovních tvarů a jejich koincidencí v textech 19. století se za předpokladu platnosti hypotézy o středověkém původu RK limitně blíží nule. Na Seydlerův článek sice reagovalo několik zastánců pravosti, diskuze se ovšem nenesla v rovině věcných argumentů, ale spíš výpadů proti využívání kvantitativních metod v lingvistice. Na dlouhou dobu tak článek zůstává jediným zástupcem rané české stylometrie.

Pomineme-li publikovanou přednášku *Použití počtu pravděpodobnosti k identifikaci textu* čerstvého absolventa Masarykovy univerzity a pozdějšího profesora Kalifornské univerzity Františka Wolfa (1928), stojí za pozornost zejména článek Romana Jakobsona *K časovým otázkám nauky o českém verši* (1935), který představuje jedno z mála – ne-li jediné využití versologie pro potřeby určování autorství v českém prostředí. Jakobson v Tomaševského šlépějích (srov. kap. 1.5) porovnává distribuce četností přízvuků a mezislovních předělů v desetislabičných verších RKZ s prokazatelně středověkými skladbami psanými desetislabičným veršem na jedné straně a s desetislabičnými verši v dílech V. Hanky, J. Lindy a dalších autorů raného obrození na straně druhé. Konstatuje přitom, že v těchto aspektech je verš RKZ mnohem bližší obrozencům než středověkým památkám. (Stejným způsobem pak analyzuje i verše osmislabičné.)

Dalším významným stylometrickým příspěvkem k problematice autorství RKZ byla rozsáhlá studie vypracovaná výzkumným týmem vedeným Marií Těšitelovou (1976). Autoři zde opět porovnávají RKZ se středověkými texty a s texty z 19. století (na rozdíl od Jakobsona s těmi, které prokazatelně vznikly před rokem 1817 a případná blízkost stylu RKZ tak nemohla být zapříčiněna vlivem rukopisů), a to na základě takových ukazatelů jako je frekvence jednotlivých slovních druhů nebo různé metriky bohatosti slovníku.

Mimo okruh RKZ je třeba zmínit především práce Pavla Vašáka věnované jak teoretickým otázkám určování autorství (1966, 1980), tak konkrétním aplikacím těchto metod, a to zejm. na texty připisované Karlu Hynku Máchovi, Josefu Barákovi (o nich podrobněji v kap. 4.2) a texty publikované pod jménem Petr Bezruč.

Jak Těšitelovou, tak Vašáka můžeme označit za reprezentanty přístupu „zlatého rysu" (srov. kap. 1.2) – jejich cílem je nalézt konkrétní izolované charakteristiky, které v rámci textů produkovaných jedním autorem zůstávají stabilní a napříč texty různých autorů vykazují co možná nejvyšší variabilitu. Metody současné stylometrie (kap. 1.3–1.4) – pomineme-li ojedinělé použití atribučních metod při jiných klasifikačních úlohách (Kubát 2016) a specifickou oblast forenzní lingvistiky (Rygl 2014, 2016) – bohužel zatím nebyly v českém prostředí systematicky testovány ani aplikovány.

9 Podrobněji o vývoji stylometrie v českém prostředí (zejm. v souvislosti s RKZ) viz Sedlačíková 2004.

1.7 Shrnutí

Na základě výše uvedeného můžeme shrnout, že od svého vzniku v 19. století vychází stylometrie z obecné představy, že autora díla lze určit na základě *podobnosti* mezi určitou *numerickou reprezentací* daného textu a numerickými reprezentacemi textů produkovaných kandidáty na jeho autorství.

Zatímco v rané stylometrii 19. století byly využívány jednoduché reprezentace jako např. délka slova měřená počtem znaků (Mendenhall), ve 20. století postupně přechází stylometrie k mnohem komplexnějším charakteristikám. Ruku v ruce s tímto vývojem se proměňovala i definice *podobnosti*: od jednoduchého (většinou optického) porovnání blízkosti dvou naměřených hodnot, přes inferenční statistiku, multidimenzionální analýzy, až po klasifikaci pomocí strojového učení.

Zmíněných charakteristik využívá současná stylometrie celou řadu: frekvence nejčetnějších slov, znakových a slovních *n*-gramů, interpunkčních znaků... Jedna podstatná součást literárního stylu (jedné podstatné literární formy) ovšem zůstává zcela mimo její pozornost. Přestože u versologických rysů se tradičně předpokládá závislost na osobě autora (např. v češtině existuje bezpočet studií na téma „čím je Máchův jamb specifický“), s jejich testováním a využitím pro potřeby atribuce básnických děl se v moderní stylometrii prakticky nesetkáme. K tomu připojme:

- Většina rysů užívaných ve stylometrii (např. frekvence slov, znakových a slovních *n*-gramů) představuje z pohledu statistiky řídké jevy, resp. přesněji jevy s distribucí LNRE (large number of rare events; srov. Baayen 2001). Ty lze smysluplně analyzovat pouze v rozsáhlých vzorcích textů (minimálně tisíce až desetitisíce slov). U básnických textů se ovšem s takovou situací v praxi setkáme zcela výjimečně – neznámé nebo zpochybněné autorství obvykle provází jednu báseň, případně několik básní, ale zřídkakdy celou básnickou sbírku. Versologické rysy na druhou stranu bývají obvykle binární (např. přízvučná/nepřízvučná slabika) nebo mohou nabývat jen poměrně malého počtu možných hodnot (např. výskyt hlásek v rýmu) a mohou tak být smysluplně analyzovány i v mnohem menších vzorcích.
- Versologické rysy lze považovat za kontextově méně závislé než obvyklé rysy užívané ve stylometrii (frekvence slov, znakových a slovních *n*-gramů). Lexikon se v rámci textů produkovaných jedním autorem bude napříč žánry a tématy nutně značně odlišovat, zatímco můžeme předpokládat, že versologické charakteristiky budou zůstávat více méně stabilní. (Dodejme, že zde není řeč o vědomě volených proměnných jako je typ metra (jamb, trochej, daktyl…) nebo strofické schéma, resp. určitá pevná forma (sonet, rondel, limerik…), ale o způsobu realizace těchto abstraktních vzorců jazykovým materiálem (viz kap. 2), který je jen těžko racionálně kontrolovatelný/napodobitelný.)
- Lexikon ve veršovaných textech závisí nejen na autorovi a žánrových/tematických charakteristikách, ale je ovlivňován i zvoleným básnickým metrem (např. Forstall a Scheirer (2010) prokázali v latinských časoměrných textech asociaci mezi metrem a frekvencemi některých znakových bigramů).
- Někteří stylometrici doporučují k dosažení spolehlivějších výsledků kombinovanou analýzu několika rysů zároveň, např. vektory určené četnostmi nejfrekventovanějších slov, znakových *n*-gramů a slovních *n*-gramů (srov. Mikros–

Perifanos 2013; Eder 2011). Tyto rysy spolu ale navzájem značně korelují. Na druhou stranu můžeme předpokládat, že versologické rysy jsou na výše jmenovaných prakticky nezávislé. Kombinovaná analýza lexikálních a versologických rysů by tak měla přinést lepší výsledky než analýza těchto rysů samostatně.

V následujících kapitolách se proto pokusíme s použitím moderních atribučních metod testovat využitelnost versologických rysů coby ukazatelů autorství, a to na česky, německy, španělsky a anglicky psaném materiálu. Takový přístup byl (dle mého nejlepšího vědomí a svědomí) zatím testován pouze v omezené míře na malých vzorcích latinské (Forstall–Jacobson–Scheirer 2011) a staroarabské poezie (Al-Falahi–Ramdani–Bellafkih 2017), v obou případech s nepříliš uspokojivými výsledky. V současné době probíhají přípravné práce na projektech, které se z tohoto pohledu zaměřují na středověkou nizozemskou poezii (Kestemont–Haverals 2018) a portugalsky psanou poezii (Mittmann–Pergher–dos Santos 2019). První výsledky našeho testování byly publikovány v Plecháč–Bobenhausen–Hammerich 2018.[10]

10 Nedlouho před dokončením této práce byl v *Nature Human Behaviour* publikován článek využívající versologické rysy pro rozlišení autorství staroanglické poezie (Neidorf et al. 2019).

2 Versologické rysy

2.1 Rytmus

Od dob ruského formalismu rozlišuje versologie mezi metrem, resp. rozměrem,[11] tj. abstraktní osnovou veršové řádky, a rytmem, tj. realizací metra konkrétními fonetickými jednotkami. Vztah mezi silnými (S) a slabými (W) pozicemi metrického vzorce a přítomností/nepřítomností určitého fonetického jevu přitom nemá deterministickou, ale statistickou povahu. Týž rozměr tak může být v různých verších rytmicky realizován různým způsobem. Příkladem budiž úvodní čtyřverší prvního zpěvu Máchova *Máje*; všechny verše jsou psány týmž rozměrem (čtyřstopý mužský jamb), ale rytmická realizace přízvučnými („1“) a nepřízvučnými slabikami („0“) je v každém verši odlišná:[12]

Byl pozdní večer – první máj –
rytmická realizace: 0 1 0 1 0 1 0 1
metrický vzorec: W_0 S_1 $W_1S_2W_2$ S_3 W_3 S_4

večerní máj – byl lásky čas.
rytmická realizace: 1 0 0 1 0 1 0 1
metrický vzorec: W_0S_1 W_1 S_2 W_2 S_3 W_3S_4

Hrdliččin zval ku lásce hlas,
rytmická realizace: 1 0 0 1 1 0 0 1
metrický vzorec: W_0S_1 W_1 S_2 W_2S_3 W_3 S_4

kde borový zaváněl háj.
rytmická realizace: 0 1 0 0 1 0 0 1
metrický vzorec: W_0 $S_1W_1S_2W_2S_3$ W_3 S_4

Předpokládá se přitom, že jednotlivé rytmické realizace nebývají v textech jednoho autora zastoupeny nahodile, ale že tvoří podstatnou součást jeho osobitého autorského stylu. Zatímco volba metra, resp. rozměru souvisí obvykle především s obecnými literárními konvencemi (např. obvyklým rozměrem české rytířské epiky 19. stol. byl čtyřstopý trochej), celkový způsob jeho rytmické realizace (rytmický styl) lze pokládat za vhodný ukazatel autorství textu.

11 Jako metrum označujeme obecně způsob alternace S-pozic a W-pozic (jamb, trochej…), jako rozměr pak varianty jednotlivých meter (pětistopý mužský jamb, čtyřstopý ženský trochej…).

12 Metrické pozice indexujeme konvenčním způsobem: index označuje pořadí daného typu pozice ve verši počínaje první S-pozicí; stojí-li tak W-pozice na samém začátku verše, je indexována nulou (zdůvodnění viz Ibrahim–Plecháč–Říha 2013: 30–31). Dodejme, že při klasifikaci slabik (1/0) vycházíme z definice R-celku (Plecháč–Kolár 2017: 46–49). Pro jednoduchost se ale přidržíme označení přízvučná/nepřízvučná.

Rytmický styl bývá ve versologii modelován dvěma hlavními metodami: (1) tzv. *rytmický profil* a (2) měření četností tzv. *rytmických typů*.[13]

2.1.1 Rytmický profil

Zavedení metody rytmického profilu bývá obvykle připisováno Andreji Bělému (1910). V české versologii se objevuje spolu s nástupem první generace pražského strukturalismu (např. Mukařovský 1934/2001), později je využíván zejm. v mnoha rozsáhlých studiích Miroslava Červenky a Květy Sgallové (např. Červenka 2006). Do dnešní doby představuje (přinejmenším v kontinentální evropské versologii) nejrozšířenější metodu rytmické analýzy.

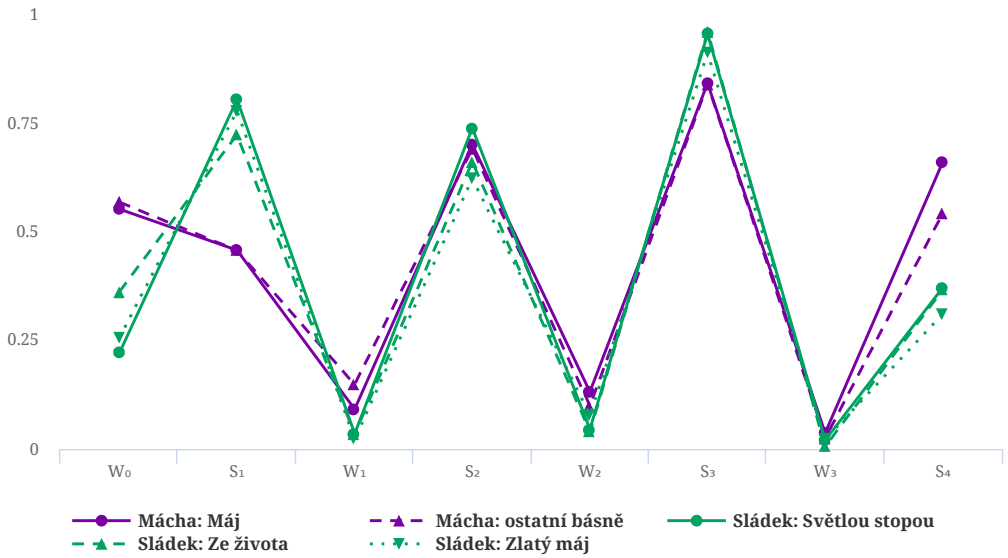
Metoda rytmického profilu spočívá v měření četností přízvuků na jednotlivých metrických pozicích verše (často pouze S-pozicích). Pro ilustraci se podívejme, jak vypadá rytmický profil všech čtyřstopých mužských jambů *Máje* a dalších Máchových veršovaných textů ve srovnání s čtyřstopými mužskými jamby obsaženými ve třech sbírkách J. V. Sládka. OBR. 2.1 ukazuje v odborné literatuře dobře zmapované rozdíly mezi rytmickými styly obou autorů:[14]

(1) Počáteční W_0-pozice vykazuje v obou Máchových souborech nápadně vyšší četnost přízvuků než v Sládkových sbírkách.

(2) S_1-pozice vykazuje v obou Máchových souborech nápadně nižší četnost přízvuků než v Sládkových sbírkách.

(3) Koncová S_4-pozice vykazuje v obou Máchových souborech nápadně vyšší četnost přízvuků než v Sládkových sbírkách

(4) Přízvuky na W_1-pozici a W_2-pozici se v obou Máchových souborech objevují o něco častěji než v Sládkových sbírkách.



OBR. 2.1: Rytmický profil čtyřstopých mužských jambů v Máchově *Máji* a ostatních básnických textech; rytmický profil čtyřstopých mužských jambů obsažených ve třech sbírkách J. V. Sládka.

13 Často se lze setkat i s označením *přízvukový profil* a *rytmické formy*.

14 Srov. Červenka 1998; Červenka–Sgallová 1978; Jirát 1931–1932; Jakobson 1938/1995.

Rytmický profil představuje jednoduchou metodu rytmické analýzy, v níž se ovšem ztrácí veškerá informace o kontextu jednotlivých slabik (srov. např. Dobritsyn 2016). Z jednotlivých hodnot zobrazených v OBR. 2.1 tak například nelze zpětně odvodit, jakou část z cca. 13 % přízvuků na W_2-pozici v Máchově *Máji* nesou přízvučná monosylaba:

	„Kde Vilém můj?“ „Viz“, plavec k ní
rytmická realizace:	0 1 0 1 1 1 0 0
metrický vzorec:	$W_0S_1W_1$ S_2 W_2 S_3W_3 S_4

a jakou víceslabičné takty:

	Kde borový zaváněl háj
rytmická realizace:	0 1 0 0 1 0 0 1
metrický vzorec:	$W_0S_1W_1S_2W_2$ S_3W_3 S_4

Zásadní problém ovšem představují tzv. extrametrické slabiky, tj. situace kdy jedné metrické pozici odpovídá více než jedna slabika.[15] Ty se sice v českém jambu a trocheji objevují zcela výjimečně, např.:

	Přistoupí strážce a lampy zář,
rytmická realizace:	1 0 0 1 0 0 1 0 1
metrický vzorec:	W_0 S_1W_1 S_2 $\llcorner W_2 \lrcorner$ S_3 W_3S_4
	(Mácha)

zejm. v anglické versifikaci se ale jedná o jev zcela běžný:[16]

	Those trackless deeps where many a weary sail
ryt. realizace:	0 1 0 1 0 1 0 0 1 0 1
metrický vzorec:	W_0 S_1 W_1 S_2 W_2 S_3 $\llcorner W_3 \lrcorner$ S_4 W_4 S_5
	(Shelley)

Stejný problém pak představují i případné nerealizované metrické pozice (⌀), v anglo-americké tradici označované jako „headless lines“:

	Znovu v mdlobách umírá
ryt. realizace:	⌀ 1 0 1 0 1 0 0
metrický vzorec:	W_0 S_1 W_1 S_2W_2 $S_3W_3S_4$
	(Mácha)

15 Interpretace dvou slabik coby realizace jedné metrické pozice (stejně jako níže zmíněné zavedení nerealizované metrické pozice) je vždy motivována kontextem. Pokud bychom např. citovaný verš *„Přistoupí strážce a lampy zář“* chtěli metricky interpretovat bez extrametrických slabik (a zároveň jej nechápat jako ojedinělý volný verš mezi ostatními jamby), museli bychom připustit, že zde (1) dochází v rámci Máchova díla k bezprecedentní situaci, kdy jsou poslední dvě S-pozice obsazeny nepřízvučnou slabikou a poslední dvě W-pozice slabikou přízvučnou (*...a*(S_3) *lam*(W_3)*-py*(S_4) *zář*(W_4)) a zároveň (2) Mácha tento ženský verš rýmuje s veršem mužským (*„Před samou věžně vstoupí tvař“*), což je opět situace, s níž se v rámci jeho díla jinde nesetkáme.

16 K důvodům odlišného postavení extrametrických slabik v české a anglické versifikaci viz Levý 1962.

Stay the King hath thrown his warder down
ryt. realizace: ⌀ 1 0 1 0 1 0 1 0 1
metrický vzorec: W_0 S_1 W_1 S_2 W_2 S_3 W_3 S_4 W_4 S_5
(Shakespeare)

Vzhledem k tomu, že metoda rytmického profilu počítá s realizací jedné metrické pozice coby s binární proměnnou (1/0), není s to takové situace zachytit.

2.1.2 Rytmické typy

Druhou základní metodu, mající kořeny rovněž v ruské škole, představuje měření četností celých rytmických realizací (rytmických typů). V českém prostředí byl tento postup uplatňován zejm. na sklonku 60. let u rozsáhlých výzkumných úkolů řešených versologickým týmem Ústavu pro českou a světovou literaturu ČSAV (např. Červenka 1971) a později v rámci mezinárodní skupiny pro výzkum slovanských versifikací (Červenka–Sgallová 1978).

Jako příklad vezměme znovu čtyřstopé mužské jamby v Máchově *Máji*, kde je všech 347 takových veršů realizováno 47 různými rytmickými typy (TAB. 2.1).

rank	rytmický typ	relativní četnost	absolutní četnost	příklad
1	10010101	0.2305	80	Večerní máj – byl lásky čas
2	01010101	0.1671	58	Byl pozdní večer – první máj
3	10010100	0.0922	32	Modré se mlhy houpají
4	01010100	0.0605	21	Já zatím hrob mu vyryji
5–6	01000100	0.0519	18	Vzdy zeleněji prosvítá
5–6	10100101	0.0519	18	Břeh je objímal kol a kol
7	01000101	0.0432	15	Tam při jezeru vížka ční
8	01001001	0.0288	10	Kde borový zaváněl háj
9–10	10011001	0.0230	8	Hrdliččin zval ku lásce hlas
9–10	10100100	0.0230	8	Dále zeleně zakvítá
			...	
31–47	100100101	0.0029	1	Přistoupí strážce a lampy zář
31–47	1010100	0.0029	1	Znovu v mdlobách umírá

TAB. 2.1: Rytmické typy čtyřstopého mužského jambu v Máchově *Máji*.

Měření frekvence rytmických typů umožňuje jednoduše zpracovat i případy, které – jak jsme viděli výše – v rytmickém profilu zpracovat nelze: extrametrické slabiky (rank 31–47, verš „Přistoupí strážce…“) a nerealizované metrické pozice (rank 31–47, verš „Znovu v mdlobách…“). Vzhledem k tomu, že rytmické typy nepracují s realizací konkrétních metrických pozic ale celých veršů, lze je využít i pro rytmickou analýzu versifikací, kde je počet takových pozic proměnlivý, resp. kde nemá smysl metrické pozice vůbec rozlišovat (tónický verš, volný verš). Na druhou stranu produkují relativně řídká data a některé autorsky specifické podřetězce se tak mohou rozdělit mezi velké množství málo frekventovaných typů.

2.1.3 Rytmické *n*-gramy

Z výše uvedených důvodů se u sylabických a sylabotónických veršů pro naše potřeby jeví jako optimální střední cesta mezi oběma popsanými metodami: měření četností rytmických realizací nikoliv celých metrických vzorců, ale jejich podřetězců (nazvěme je *rytmické n-gramy*). U veršů o k metrických pozicích tak budeme měřit nikoliv četnosti realizací celého řetězce pozic, ale četnosti realizací všech podřetězců o n pozicích začínajících na i-té slabice, kde $i \in \{1, 2, 3, \ldots, k - n\}$. Postup ilustruje TAB. 2.2 uvádějící četnosti rytmických bigramů v Máchově *Máji*.

	realizace									
	00	**01**	**10**	**11**	**000**	**001**	**011**	**101**	**100**	**ø1**
$\mathbf{W_0S_1}$	0	0.4092	0.5533	0.0346	0	0	0	0	0	0.0029
$\mathbf{S_1W_1}$	0.4611	0.0893	0.4467	0	0	0.0029	0	0	0	0
$\mathbf{W_1S_2}$	0.2017	0.7061	0.0836	0.0058	0	0	0.0029	0	0	0
$\mathbf{S_2W_2}$	0.2104	0.0749	0.6340	0.0490	0	0	0	0.0029	0.0029	0
$\mathbf{W_2S_3}$	0.0432	0.8012	0.1066	0.017	0	0.0029	0.0029	0	0	0
$\mathbf{S_3W_3}$	0.1239	0.0259	0.8242	0.0086	0	0	0	0.0029	0.0144	0
$\mathbf{W_3S_4}$	0.3083	0.6397	0.0345	0	0.0058	0.0086	0.0029	0	0	0

TAB. 2.2: Rytmické bigramy v čtyřstopých mužských jambech Máchova *Máje*.

Abychom zachytili co možná nejpestřejší škálu rytmických specifik, budeme při atribučních experimentech reprezentovat jednotlivé vzorky kombinací četností rytmických bigramů, trigramů a tetragramů.

2.2 Rým

Kromě rytmického stylu bývá za autorsky specifický považován i styl rýmování. Pro naše potřeby budeme rým reprezentovat jako neuspořádanou dvojici následujících charakteristik obou rýmových slov:

(1) morfologické rysy (v češtině slovní druh určený první pozicí morfologické značky, v ostatních jazycích celá značka produkovaná stochastickým taggerem TreeTagger),[17]

(2) délka slova měřená počtem slabik,

(3) počet slabik za poslední přízvučnou slabikou,

(4) koda poslední slabiky,

(5) jádro poslední slabiky,

(6) pretura poslední slabiky + koda předposlední slabiky (pouze ženské rýmy),

(7) jádro předposlední slabiky (pouze ženské rýmy).

TAB. 2.3 ukazuje takovou reprezentaci tří vybraných rýmů z prvního zpěvu Máchova *Máje*.

Textové vzorky reprezentujeme relativními četnostmi jednotlivých neuspořádaných dvojic v rámci dané kategorie.

17 Výpis příslušných tagsetů viz <http://www.cis.uni-muenchen.de/~schmid/tools/ TreeTagger/>.

	máj : háj	bledá : hledá	vlnou : plnou
(1)	{N, N}	{A, V}	{N, A}
(2)	{1, 1}	{2, 2}	{2, 2}
(3)	{0, 0}	{1, 1}	{1, 1}
(4)	{j, j}	{ø, ø}	{ø, ø}
(5)	{a:, a:}	{a:, a:}	{o͡u, o͡u}
(6)	neurčuje se	{d, d}	{n, n}
(7)	neurčuje se	{e, e}	{l̩, l̩}

TAB. 2.3: Reprezentace vybraných rýmů Máchova *Máje* podle výše uvedených kategorií (česká fonetická traskripce).

2.3 Eufonie

Za autorsky specifický rys bývá konečně považována i tzv. hlásková instrumentace, tj. kumulace týchž nebo podobných hlásek, resp. hláskových shluků překračující jazykovou pravděpodobnost (srov. Červenka 2002), která bývá dále rozdělována na eufonii („příjemné“) a kakofonii („nepříjemné“).

Už v 60. letech navrhnul Gabriel Altmann (1966a, 1966b) způsob, jak přítomnost hláskové instrumentace, resp. jednoho ze způsobů její realizace testovat. (Altmann používá pro hláskovou instrumentaci metonymicky termín „eufonie“ – pro jednoduchost se tohoto označení přidržíme.) Tato metoda je v posledních letech využívána při analýze konkrétních básnických textů (Wimmer et al. 2003, Čech–Popescu–Altmann 2011, 2014) i analýze vývojových tendencí (Popescu et al. 2015). Vzhledem k tomu, že sami autoři tuto metodu zmiňují jako možný indikátor autorství básnických textů (Čech–Popescu–Altmann 2014: 110), podíváme se na ni podrobněji.

Jedná se o přímočarou aplikaci binomického testu, kdy jsou hlásky každého verše nejprve rozděleny na dvě třídy: vokalické (V) a konsonantické (K) a následně se pro každou hlásku, která se ve verši vyskytuje alespoň dvakrát, řeší úloha, zda lze její frekvenci v rámci všech hlásek dané třídy ve verši považovat s ohledem na nějaký referenční korpus za statisticky významnou.

Mějme tedy hlásku s náležející do jedné ze dvou tříd $c(s) \in \{V, K\}$, která se v konkrétním verši vyskytuje celkem $m \geq 2$ krát, přičemž hlásek třídy $c(s)$ je v tomto verši celkem n. Dále mějme relativní četnost f hlásky s mezi všemi hláskami náležejícími do třídy $c(s)$ v referenčním korpusu. Pravděpodobnost, že opakování hlásky s je ve verši dílem náhody, pak lze spočíst jako

$$p = \sum_{x=m}^{n} \binom{n}{x} f^x (1-f)^{n-x} \qquad (2.1)$$

Platí-li $p < \alpha$ (tj. pravděpodobnost je nižší než hladina významnosti konvenčně stanovovaná na $\alpha = 0{,}05$), je eufonický koeficient hlásky s v daném verši vypočten jako

$$E(s) = 100(\alpha - p) \qquad (2.2)$$

Je-li $p \geq \alpha$, pak:

$$E(s) = 0 \qquad (2.3)$$

Eufonický koeficient každého verše je pak vypočten jako aritmetický průměr nenulových koeficientů v něm obsažených opakování hlásek, eufonický koeficient básně jako průměr eufonických koeficientů jednotlivých veršů.

Proti takovému přístupu lze ale vnést několik námitek:

(1) Některé hlásky – jako např. dlouhé a krátké varianty týchž samohlásek – bývají při analýzách počítány za jeden typ, což je pochopitelné. Problém ale nastává u slabikotvorných a neslabikotvorných variant téhož konsonantu. Pokládají-li např. Čech–Popescu–Altmann (2014) v češtině slabikotvorné [r̩] i neslabikotvorné [r] za jeden typ (totéž v případě [l̩]/[l]), dochází tím u veršů, které obsahují x slabikotvorných konsonantů, k posunu $n_K + x$ a $n_V - x$. Verš je tedy modelován jako sekvence dvou typů „slotů“ (V, K) a máme dvojice percepčně i artikulačně velice podobných hlásek ([r=]/[r], [l=]/[l]), které ovšem obsazují „sloty“ různé. Tento problém je v altmannovském přístupu řešen jednostranně a bez patřičného teoretického zdůvodnění.

(2) Výše uvedená definice eufonického koeficientu je s ohledem na to, co má modelovat, zavádějící. Vezměme modelový příklad dvou veršů a předpokládejme, že vyznačená opakování byla vyhodnocena jako statisticky významná a byly jim přiděleny následující koeficienty:

(a) Kost**e**lní **hl**a**h**o**l** zval **h**ora**l**y
E(l) = 4 *E*(h) = 1 *E*(verš) = (4 + 1) / 2 = 1,5

(b) Kost**e**lník **L**e**l**ek zva**l** hora**l**y
E(l) = 4 *E*(verš) = 4

Oba verše obsahují stejný počet vokalických (=9) i konsonantických hlásek (=15) a pět výskytů hlásky [l]. První verš krom toho obsahuje i (méně významné) opakování hlásky [h]. Celkové eufonické koeficienty veršů ale odpovídají paradoxnímu předpokladu, že eufonie prvního verše je opakováním hlásky [h] nějakým způsobem umenšována; *E*(verš 1) < *E*(verš 2). Eufonický koeficient by proto bylo vhodnější modelovat ne jako aritmetický průměr, ale např. jako součet. Tím by se ale koeficienty různě dlouhých veršů staly navzájem neporovnatelné a bylo by třeba přikročit ještě k nějaké normalizaci.

(3) Předpoklad nezávislosti jevů představuje zásadní slabinu Altmannovy metody. Binomické rozdělení popisuje pravděpodobnost výskytu jevu v *nezávislých* pokusech. Modelovým příkladem je situace, kdy máme urnu, v níž je m modrých a n červených míčků, a ptáme se, jaká je pravděpodobnost, že vytáhneme-li najednou x míčku, bude alespoň y z nich červených. Pravděpodobnosti se ale znatelně změní ve chvíli, kdy do většiny červených míčků umístíme magnet, tzn. budou mít tendenci v urně vytvářet shluky. Pravděpodobnost souvýskytu červených míčků ve výběru tak už nebude odpovídat součinu jejich relativních četností v urně – vytáhneme-li alespoň jeden červený, je vysoká pravděpodobnost, že jich spolu s ním vytáhneme víc. A právě to je velice častá situace v jazyce. Příkladem je v češtině diftong [o̯ʊ̯], který sice v rámci vokalických hlásek patří k málo frekventovaným, má ale vysokou pravděpodobnost souvýskytu. (To je dáno zejm. jeho výskytem v koncovkách, např. *s krásnou novou knihou*). Altmannova metoda má tak tendenci klasifikovat jako projev eufonie i konfigurace, které jsou v jazyce zcela běžné. Jinak řečeno, je náchylná k chybám prvního typu.

Z výše uvedených důvodů proto Altmannovu metodu necháme stranou a „eufoničnost“ básně budeme aproximovat prostými četnostmi výskytu jednotlivých hlásek.

3 Experimenty

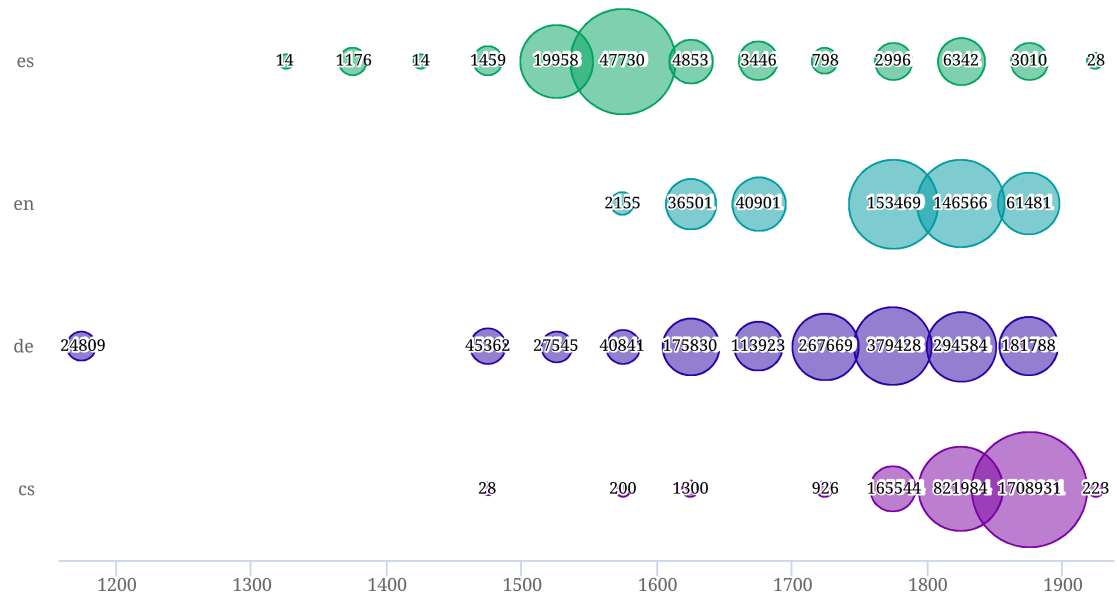
3.1 Data

Atribuční síla versologických rysů byla testována na čtyřech korpusech básnických textů: *Korpus českého verše* (Plecháč 2016; Plecháč–Kolár 2015), Korpus německého verše *Metricalizer* (Bobenhausen–Hammerich 2015; Bobenhausen 2011), španělský *Corpus de Sonetos del Siglo de Oro* (Navarro-Colorado–Ribes-Lafoz–Sánchez 2016; Navarro-Colorado 2015) a *Guttenberg English Poetry Corpus* (Jacobs 2018). Pro jednoduchost je budeme nadále označovat jako CS, DE, ES a EN.

Rámcovou charakteristiku korpusů podává TAB. 3.1 a OBR. 3.1.

	# autorů	# básní	# veršů	# tokenů
CS	613	80 229	2 727 632	14 923 528
DE	248	53 608	1 716 348	10 462 211
ES	52	5078	71 150	465 982
EN	46	10 079	441 073	3 193 989

TAB. 3.1: Velikost korpusů.



OBR. 3.1: Počty veršů v jednotlivých korpusech podle data narození autora (interval 50 let). Velikost bubliny znázorňuje relativní velikost v daném korpusu (měřenou počtem veršů).

Je zřejmé, že atribuční experimenty vyžadují korpusy důkladně proznačkované. Jak ukazuje TAB. 3.2, ve výchozím stavu obsahoval ale všechny potřebné úrovně značkování pouze korpus CS.

	CS	DE	ES	EN
tokenizovaný	1	1	0	0
lemmatizovaný	1	0	0	0
morfologicky označkovaný	1	0	0	0
foneticky transkribovaný	1	1	0	0
anotace meter	1	1	–	0
anotace přízvuků	1	1	1	0
anotace rýmů	1	1	0	0

TAB. 3.2: Výchozí stav proznačkování jednotlivých korpusů (1: označkováno, 0: neoznačkováno, –: nerozlišuje se).

Pro potřeby této práce bylo tedy provedeno dodatečné značkování. Tokenizace (ES, EN), lemmatizace (DE, ES, EN) a morfologické značkování (DE, ES, EN) bylo provedeno pomocí stochastického taggeru *TreeTagger* (Schmid 1994), fonetická transkripce (ES, EN) a anotace přízvuků (EN) byla provedena pomocí populárního TTS syntetizéru *Espeak*, anotace meter (EN) byla provedena pomocí balíčku *Prosodic* (Antilla–Heuser 2016), anotace rýmů (ES, EN) byla provedena pomocí balíčku *RhymeTagger* (Plecháč 2018).

3.1.1 Přesnost značkování

Je zřejmé, že klíčovým faktorem jakéhokoliv automatického značkování je jeho úspěšnost. Pro většinu nástrojů použitých pro dodatečné značkování i nástrojů použitých autory korpusů byly publikovány empirické odhady:

Morfologické značkování, lemmatizace, tokenizace

- Horsmann, Erbs a Zesch (2015) testovali úspěšnost morfologického značkování pomocí *TreeTaggeru* na **anglických** datech z *British National Corpus*, *Brown Corpus* a *GUM*. U psaných textů uvádějí úspěšnost **0,94** (četnost tokenů se správně přidělenou značkou).
- V tomtéž článku uvádějí Horsmann, Erbs a Zechs výsledky evaluace morfologického značkování pomocí *TreeTaggeru* na **německých** datech z korpusu *Tüba-DZ*. Giesbrechtová a Evert (2009) provedli testování na rozsáhlém korpusu novinových článků *TIGER*, validaci na relativně malém vzorku provedl i sám autor *TreeTaggeru* (Schmid 1994). Ve všech případech jsou uváděny hodnoty okolo **0,97** (četnost tokenů se správně přidělenou značkou).
- Göhringová (2009) provedla u *TreeTaggeru* evaluaci morfologického značkování na souboru 200 manuálně označkovaných **španělských** vět. Na rozdíl od výše uvedených nepracovala s úspěšností značkování na počet tokenů, ale měřila hodnoty *precision* a *recall* jednotlivých značek z tagsetu. Pro obé uvádí mikroprůměr **0,94**.
- Spoustová et al. (2007) a Skoumalová (2011) provedli evaluaci kombinovaného stochasticko-pravidlového taggeru použitého při značkování korpusu **CS** na testovacím subkorpusu PDT s výsledkem **0,95** (četnost tokenů se správně přidělenou značkou).

- Vzhledem k tomu, že lemmatizace je u obou zmíněných taggerů svázána s morfologickým značkováním (desambiguace), můžeme uvedené hodnoty vztáhnout i k ní. Totéž platí i pro rovinu tokenizace.

Anotace meter a přízvuků

- Heuser (github.com/quadrismegistus/prosodic) vytvořil testovací vzorek sestávající z 1800 **anglicky** psaných veršů. U každého verše byl dvěma subjekty manuálně anotován metrický vzorec. Četnost veršů, u nichž oba anotátoři zanesli shodný vzorec činila 0,94–0,99 v závislosti na typu metra. Míra shody s anotací pomocí balíčku *Prosodic* činila **0,93** pro jamb, **0,95** pro trochej, **0,85** pro anapest a **0,84** pro daktyl.
- Odhad úspěšnosti anotace meter/přízvuků v korpusu **DE** nebyl publikován.
- Navarro-Colorado (2017) vytvořil testovací vzorek sestávající ze 100 sonetů obsažených v korpusu **ES**. Vzorek byl manuálně rytmicky anotován třemi subjekty. Četnost veršů, u nichž všichni tři anotátoři zanesli shodné schéma, činila 0,96. Míra shody automatické rytmické anotace v korpusu ES s alespoň dvěma anotátory pak činila **0,95**.
- U korpusu **CS** byla na základě manuálně označkovaných vzorků zjištěna úspěšnost rozpoznávání meter jednotlivých veršů 0,95 (Plecháč 2016).

Anotace rýmů

- Úspěšnost balíčku *RhymeTagger* byla testována na anglicky, francouzsky a česky psané poezii (Plecháč 2018). Na manuálně označkovaných datech byly zjištěny následující hodnoty *precision* (P) a *recall* (R): **EN: P = 0,96; R = 0,88; FR: P = 0,94; R = 0,87; CS: P = 0,94; R = 0,96**.

Výše uvedené výsledky dávají poměrně optimistický obraz kvality značkování v našich korpusech. Na druhou stranu je třeba dodat, že se evaluace jednotlivých rovin napříč korpusy metodologicky značně odlišují a že u lingvistického značkování našich dat můžeme téměř bezpečně předpokládat nižší než uváděnou úspěšnost danou jednak specifiky básnického jazyka (novotvary, slovosledné inverze…), jednak starším datem vzniku textů. Z těchto důvodů jsme provedli vlastní (méně rozsáhlé) sondy.

Odhad přesnosti značkování v jednotlivých korpusech byl proveden na základě manuální kontroly náhodných vzorků rodilými mluvčími s filologickým vzděláním.[18] Vzhledem k tomu, že značkované jevy se týkají jednotek z různých jazykových rovin (od jednotlivých hlásek po rýmy spojující často různé větné celky), byly pro každý subkorpus vytvořeny tři sady různě rozsáhlých vzorků:

(1) vzorek pro validaci tokenizace, lemmatizace, morfologického značkování, fonetické transkripce a detekce přízvuků (CS: 52 veršů / 287 tokenů / 511 slabik, DE: 55 veršů / 244 tokenů / 377 slabik, ES: 98 veršů / 627 tokenů / 1078 slabik, EN: 54 veršů / 346 slov / 436 slabik); v každém vzorku CS, DE a EN byly obsaženy

18 Za ochotnou pomoc srdečně děkuji následujícím kolegům: **CS:** Michal Kosák (Ústav pro českou literaturu AV ČR); **DE:** Michael Wögerbauer (Ústav pro českou literaturu AV ČR); **ES:** Helena Bermúdez-Sabel (Université de Lausanne) a Clara Isabel Martínez Cantón (Universidad Nacional de Educación a Distancia, Madrid); **EN:** David Birnbaum (University of Pittsburgh).

počáteční pasáže[19] náhodně vybraných básní čítající vždy alespoň 8 veršů, vzorek ES sestával ze 7 náhodně vybraných sonetů;

(2) vzorek pro validaci značkování veršových rozměrů (CS: 120 veršů, DE: 114 veršů, EN: 118 veršů); v každém vzorku byly obsaženy počáteční pasáže náhodně vybraných básní čítající vždy alespoň 8 veršů;

(3) vzorek pro validaci detekce rýmů (CS: 86 rýmů, DE: 97 rýmů, ES: 183 rýmů, EN: 84 rýmů); v každém vzorku CS, DE a EN byly obsaženy počáteční pasáže náhodně vybraných básní tak, aby žádný rým nepřesahoval mimo vybrané pasáže, vzorek ES sestával z 20 náhodně vybraných sonetů.

TAB. 3.3 a TAB. 3.4 udávají pro jednotlivé značkované jevy relativní zastoupení značek, které byly validátory označeny jako správně přiřazené (úspěšnost). U detekce rýmů uvádíme jak hodnotu precision (jak velká část označkovaných rýmů představuje skutečný rým), tak hodnotu recall (jak velká část rýmů ve vzorku byla označkována). Vzhledem k tomu, že morfologické značkování využíváme pouze v případě rýmových slov (srov. kap. 2.2), uvádíme kromě celkové úspěšnosti i úspěšnost pouze pro koncová slova veršů.

			morfologické značkování		fonetická transkripce
	tokenizace[20]	lemmatizace	vše	koncová slova	
CS	1	0,9692	0,9577	0,9302	1
DE	1	0,9385	0,9590	0,9836	1
ES	1	0,9426	0,9011	0,9984	0,9936
EN	0,9844	0,9855	0,9422	0,9259	0,9595

TAB. 3.3: Odhad úspěšnosti tokenizace, lemmatizace, morfologického značkování a fonetické transkripce v jednotlivých korpusech.

	detekce rýmů		detekce přízvuků	detekce rozměrů
	precision	recall		
CS	0,9882	0,9767	1	1
DE	1	0,9794	0,9602	1
ES	0,9800	1	0,9944	-
EN	1	0,9231	0,9679	0,8833

TAB. 3.4: Odhad úspěšnosti detekce rýmů, přízvuků a veršových rozměrů v jednotlivých korpusech.

Zjištěné hodnoty ukazují, že všechny roviny značkování lze u všech korpusů považovat za poměrně spolehlivé. Jediný případ, kdy byla zjištěna úspěšnost nižší než 0,9, představuje detekce rozměrů ve vzorku anglické poezie (~0,88). Dodejme ale, že většinu chybných anotací zde představují případy, kdy je pro správnou detekci rozměru třeba počítat s tím, že se běžná výslovnost přizpůsobuje veršovému rytmu (např. *generous* ~ [dʒɛn.rəs]), což lze pokládat za jev značně subjektivní, a že se všechny chyby týkají pouze délky rozměru (počet stop, resp. typ klauzule) a nikoliv rozpoznání metra samotného.

19 Počáteční pasáže byly vybrány z toho důvodu, aby validátoři měli k dispozici dostatečný kontext ke správnému posouzení desambiguace. Totéž platí i pro validaci veršových rozměrů (např. případná změna metra uvnitř vzorku může mít za následek chybnou interpretaci validátorem) a validaci detekce rýmů (chybná interpretace ve chvíli, kdy jeden člen rýmového páru leží mimo vzorek).

20 Ve vzorku z korpusu EN představují všechny chyby historické varianty zápisu préterit indikující, že sufix *-ed* je pokládán za neslabičný – např. *listen'd* namísto *listened* – které byly chybně interpretovány jako *listen* + *would*.

3.1.2 Subkorpusy

Z výše popsaných korpusů bylo extrahováno 10 subkorpusů (CS1, CS2, CS3, DE1, DE2, DE3, ES1, ES2, EN1, EN2) sdružujících autory narozené v určitém časovém rozmezí. Při stanovení těchto rozmezí jsme se řídili jednak požadavkem na dostatečný objem dat (viz níže), jednak jsme se snažili rámcově vycházet z obvyklé literárněhistorické periodizace (např. CS1 obsahuje autory „obrozenecké“, CS2 obsahuje převážně autory spojované s „lumírovskou“, resp. „ruchovskou“ školou, EN1 obsahuje kanonické představitele anglického romantismu).

Pro každý subkorpus byl vybrán jeden typický veršový rozměr, resp. skupina rozměrů (CS1: kombinace čtyřstopého mužského a ženského trocheje (T4), CS2–3: pětistopý ženský jamb (I5f), DE1–3: tónický verš (F), ES1–2: 11slabičný sylabický verš (11σ),[21] EN1–2: pětistopý mužský jamb (I5m)).

Autoři v jednotlivých subkorpusech jsou reprezentováni vzorky svých básní v příslušném rozměru / příslušných rozměrech. Každý vzorek čítá 100 veršů s minimálně 40 rýmovými páry, přičemž v jednom vzorku může být obsaženo více básní, žádná báseň ale není rozdělena do více vzorků. U každého autora bylo požadováno minimálně 10 takových vzorků.

Podrobnější informace o jednotlivých subkorpusech podává TAB. 3.5.

3.2 Atribuce na základě versologických rysů

Při první sadě experimentů byla testována samotná úspěšnost atribuce na základě versologických rysů.

Každý z výše uvedených 10 subkorpusů byl zredukován na 50 vzorků, a to dvojím náhodným výběrem: (1) náhodným výběrem 5 autorů (v případě CS3, ES1 a EN1–2 byli tímto vybráni všichni autoři) a (2) náhodným výběrem 10 vzorků od každého autora. Každý vzorek byl reprezentován vektorem určeným versologickými rysy, které byly podrobně popsány v kap. 2:

(1) četnostmi rytmických 2-, 3- a 4-gramů u sylabických a sylabotónických veršů (CS, ES, EN); četnostmi 100 nejfrekventovanějších rytmických typů u tónických veršů (DE);

(2) četnostmi morfologických, fonetických a rytmických charakteristik rýmů;

(3) četnostmi jednotlivých hlásek.

Jako klasifikátor byl zvolen Support Vector Machine s klasifikací do více tříd strategií *one-vs.-one* (srov. kap. 1.4.2) implementovaný v modulu *SVC* knihovny *scikit-learn*[22] s následujícím nastavením (srov. kap. 1.4.1):

- *kernel='linear'* (Klasifikátor bez jádrové transformace dat.)
- *C=1* (Defaultní hodnota penalizačního parametru. I při testování jiných nastavení zůstávaly výsledky poměrně stabilní.)

Ostatní volitelné parametry modulu byly ponechány v defaultním nastavení.

21 Jediný rozměr obsažený v korpusu ES.

22 <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

zkratka	rozměr(y)	narozeni	# autorů	autoři (# vzorků)
CS1	T4	1760–1820	9	Čelakovský, František Ladislav (12); Havelka, Matěj (13); Hněvkovský, Šebestián (11); Kulda, Beneš Metod (27); Nejedlý, Vojtěch (17); Picek, Václav Jaromír (21); Pohan, Václav Alexander (10); Tablic, Bohuslav (16); Vinařický, Karel Alois (15);
CS2	I5f	1840–1855	7	Čech, Svatopluk (13); Kvapil, František (11); Mokrý, Otokar (15); Nečas, Jan Evangelista (10); Sládek, Josef Václav (16); H. Uden (17); Vrchlický, Jaroslav (281)
CS3	I5f	1860–1870	5	Klášterský, Antonín (64); Kvapil, Jaroslav (19); Leubner, František (10); Machar, Josef Svatopluk (22); Sova, Antonín (15)
DE1	F	1650–1699	6	Brockes, Barthold Heinrich (51); Drollinger, Carl Friedrich (11); Gottsched, Johann Christoph (29); Kuhlmann, Quirinus (30); Neukirch, Benjamin (21); Tersteegen, Gerhard (25)
DE2	F	1730–1754	5	Goethe, Johann Wolfgang (46); Jacobi, Johann Georg (12); Müller, Friedrich (15); Pfeffel, Gottlieb Konrad (28); Wieland, Christoph Martin (23)
DE3	F	1760–1794	7	Bernhardi, Sophie (12); Eichendorff, Joseph von (32); Grillparzer, Franz (52); Müller, Wilhelm (16); Schenkendorf, Max von (10); Schulze, Ernst (19); Tieck, Ludwig (28)
ES1	11σ	1500–1560	5	de Acunya, Hernando (10); de Borja, Francisco (17); de Cetina, Gutierre (31); de Góngora, Luis (14); de Herrera, Fernando (39);
ES2	11σ	1561–1599	6	Argensola, Bartolome (19); de Quevedo, Francisco (63); de Rojas, Pedro Soto (15); de Tassis y Peralta, Juan (25); de Ulloa y Pereira, Luis (13); de Vega, Lope (167);
EN1	I5m	1750–1799	5	Byron, George Gordon (46); Coleridge, Samuel Taylor (27); Keats, John (15); Shelley, Percy Bysshe (31); Wordsworth, William (23)
EN2	I5m	1800–1869	5	Bierce, Ambrose (20); Browning, Elizabeth Barrett (11); Hardy, Thomas (17); Lowell, James Russell (16); Wilde, Oscar (15);

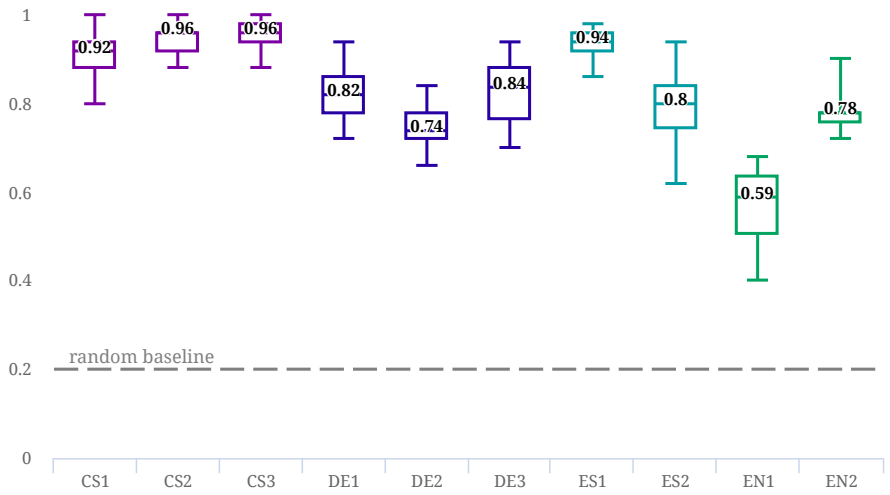
TAB. 3.5: Detaily vytvořených subkorpusů (T4: kombinace ženského a mužského čtyřstopého trocheje, I5f: pětistopý ženský jamb, I5m: pětistopý mužský jamb, F: tónický verš, 11σ: 11slabičný sylabický verš).

Pro každý subkorpus byl proveden odhad úspěšnosti modelu *leave-one-out* křížovou validací.[23] Abychom získali reprezentativnější výsledky, byl výše popsaný proces 30krát zopakován, v každé iteraci vždy s novým náhodným výběrem autorů i jejich vzorků.

Výsledky křížových validací (OBR. 3.2)[24] ukazují, že versologické rysy lze považovat za smysluplný stylometrický ukazatel – každá z 300 hodnot několikanásobně převyšuje

23 Vzhledem k tomu, že pracujeme s relativně malými počty vzorků na jednu třídu, výsledek běžné *leave-one-out* validace by teoreticky mohl být negativně ovlivněn menším zastoupením vzorků ze správné třídy v trénovacích datech (u klasifikovaného vzorku je skutečný autor zastoupen 9 vzorky, zatímco ostatní autoři jsou zastoupeni 10 vzorky). Abychom toto riziko eliminovali, vyřazujeme při každé klasifikaci z trénovacích dat náhodně jeden vzorek od každého autora s výjimkou autora klasifikovaného vzorku. Tento postup je uplatněn u všech experimentů popsaných v kapitole 3.

24 Veškeré boxploty v této práci jsou konstruovány následovně: vnitřní hradba zobrazuje mezikvartilové rozpětí ($Q_{0,25}$; $Q_{0,75}$), linie mezi nimi zobrazuje medián ($Q_{0,5}$), jehož zaokrouhlená hodnota je uvedena v popisku. Vnější hradba zobrazuje minimum a maximum distribuce.



OBR. 3.2: Výsledky křížových validací modelů založených na versologických rysech (30 iterací s náhodným výběrem vzorků).

hranici *random baseline* (při 5 autorech zastoupených vždy 10 vzorky RB = $\sum_{i=1}^{5}\left(\frac{10}{50}\right)^2$ = 0,2).[25] Mezi výsledky jednotlivých subkorpusů ale můžeme sledovat poměrně vysokou variabilitu. Poměrně zřetelně můžeme vyčlenit tři třídy (shluky):

(1) *Vysoce úspěšné modely*, (CS1–3, ES1), kde se medián hodnot pohybuje v rozmezí 0,92–0,94 a dolní kvartil v rozmezí 0,88–0,94.

(2) *Úspěšné modely* (DE1–3, ES2, EN2), kde se medián hodnot pohybuje v rozmezí 0,74–0,84 a dolní kvartil v rozmezí 0,72–0,78.

(3) *Neúspěšný model* (EN1), kde medián hodnot činí pouze 0,59 a dolní kvartil 0,51.

Důvody celkové variability výsledků mohou být různé a vzhledem k tomu, že strojové učení má obecně povahu černé skříňky (tzn. známe vstup a výstup, ale máme jen málo informací o tom, co se děje uvnitř), jen těžko prokazatelné.

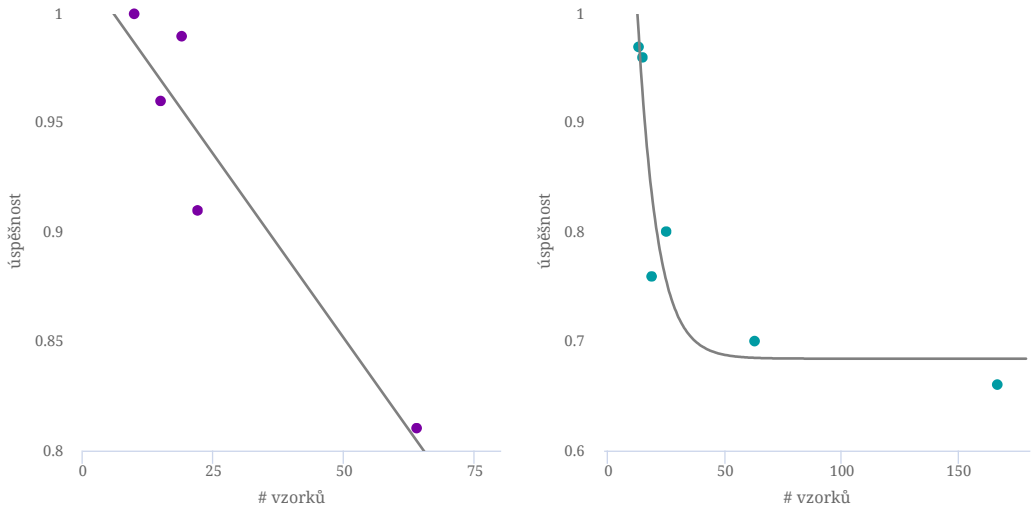
3.2.1 Předpoklady rozpoznatelnosti autorství

Zdá se, že nezanedbatelnou roli hraje u jednotlivých autorů celkový objem dat, z nichž je prováděn náhodný výběr. Autoři s vysokým počtem vzorků, jako je např. Jaroslav Vrchlický (281 vzorků), Lope de Vega (167), Francisco de Quevedo (63), Johann Wolfgang Goethe (46), Barthold Heinrich Brockes (51) nebo Franz Grillparzer (52) vykazují zpravidla nižší úspěšnost než ostatní autoři v daném subkorpusu (viz TAB. 3.6). U některých subkorpusů má dokonce vztah mezi počtem vzorků a úspěšností povahu funkční závislosti (OBR. 3.3). Budeme-li předpokládat, že objemnější dílo s sebou nese i větší stylovou variabilitu, jde o jev vcelku intuitivní.

25 Kromě SVM byly testovány i další klasifikátory: random forest (implementace v modulu ensemble knihovny scikit-learn; <https://scikit-learn.org/stable/modules/classes.html#module-sklearn.ensemble>), naïve Bayes classifier (implementace v modulu naive_bayes knihovny scikit-learn; <https://scikit-learn.org/stable/modules/classes.html#module-sklearn.naive_bayes>) a míry z rodiny Delta (vlastní implementace v jazyce Python). Ve všech případech ale vykazovaly znatelně nižší úspěšnost než SVM.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
CS1	**(1) Čelakovský**	**0,8**	0,07					0,06		
	(2) Havelka	0,1	**0,84**					0,03		0,03
	(3) Hněvkovský	0.02		**0,95**	0,01	0,02		0,01	0,04	
	(4) Kulda				**0,99**				0,01	
	(5) Nejedlý			0,03	0,01	**0,98**			0,09	
	(6) Picek	0,04					**1**	0,02		
	(7) Pohan	0,05	0,08					**0,84**		0,03
	(8) Tablic			0,02					**0,86**	
	(9) Vinařický		0,01					0,04		**0,94**
CS2	**(1) Čech**	**1**	0,06				0,03	0,03		
	(2) Kvapil		**0,92**					0,04		
	(3) Mokrý			**1**						
	(4) Nečas				**1**	0,03				
	(5) Sládek					**0,9**		0,03		
	(6) Uden					0,06	**0,96**	0,04		
	(7) Vrchlický		0,02				0,01	**0,86**		
CS3	**(1) Klášterský**	**0,91**	0,02							
	(2) Kvapil	0,06	**0,98**							
	(3) Leubner	0,01		**1**						
	(4) Machar	0,01			**0,95**	0,03				
	(5) Sova				0,05	**0,97**				
ES1	**(1) Acunya**	**0,93**		0,07						
	(2) Borja		**1**							
	(3) Cetina	0,04		**0,91**	0,01					
	(4) Góngora			0,02	**0,86**	0,02				
	(5) Herrera	0,04		0,01	0,13	**0,98**				
ES2	**(1) Argensola**	**0,73**	0,04	0,03	0,07		0,12			
	(2) Quevedo	0,08	**0,66**		0,16		0,06			
	(3) Rojas			**0,93**			0,08			
	(4) Tassis y P.	0,09	0,24		**0,76**		0,03			
	(5) Ulloa y P.	0,03	0,02			**1**				
	(6) Vega	0,08	0,03	0,04	0,01		**0,71**			
EN1	**(1) Byron**	**0,66**	0,06	0,03	0,03	0,15				
	(2) Coleridge	0,09	**0,56**	0,14	0,05	0,14				
	(3) Keats	0,01	0,13	**0,63**	0,15	0,05				
	(4) Shelley	0,06	0,03	0,08	**0,53**	0,19				
	(5) Wordsworth	0,18	0,22	0,12	0,24	**0,47**				
EN2	**(1) Bierce**	**0,99**	0,04	0	0,01					
	(2) Browning	0,01	**0,59**	0,18	0,03					
	(3) Hardy		0,09	**0,66**	0,04					
	(4) Lowell		0,15	0,08	**0,65**					
	(5) Wilde		0,13	0,08	0,27	**1**				
DE1	**(1) Brockes**	**0,76**	0,06	0,1	0,04	0,05	0,09			
	(2) Drollinger	0,07	**0,84**	0,11	0,05	0,02				
	(3) Gottsched	0,17	0,1	**0,77**	0,03	0,1	0,06			
	(4) Kuhlmann				**0,88**					
	(5) Neukirch			0,01		**0,83**				
	(6) Tersteegen			0,01			**0,85**			
DE2	**(1) Goethe**	**0,53**		0,07	0,01	0,05				
	(2) Jacobi	0,22	**0,83**	0,1	0,04	0,02				
	(3) Müller	0,19	0,03	**0,76**	0,04	0,09				
	(4) Pfeffel	0,03	0,13	0,01	**0,85**	0,08				
	(5) Wieland	0,03		0,05	0,06	**0,76**				
DE3	**(1) Bernhardi**	**0,97**								
	(2) Eichendorff		**0,88**	0,1	0,1	0,03	0,05	0,03		
	(3) Grillparzer			**0,7**						
	(4) Müller		0,01	0,03	**0,72**	0,02	0,04	0,03		
	(5) Schenkendorf		0,05	0,08	0,1	**0,89**	0,02	0,05		
	(6) Schulze	0,02	0,01	0,04	0,01		**0,82**	0,1		
	(7) Tieck	0,01	0,06	0,05	0,07	0,06	0,08	**0,79**		

TAB. 3.6: Chybové matice modelů založených na versologických rysech (relativní četnosti). Řádky udávají autora predikovaného modelem, sloupce udávají autora skutečného (jemu přidělené číslo v popisech řádků). Jednotlivé hodnoty udávají s jakou relativní četností se daná predikce u vzorků daného autora objevila.



OBR. 3.3: Vztah mezi úspěšností rozpoznávání autora a počtem jeho vzorků. Vlevo Subkorpus CS3: aproximace funkcí $y = 0{,}9991 - 0{,}0015x$ ($R^2 = 0.86$); vpravo subkorpus ES2: $y = 0{,}693 + 5{,}395e^{-0{,}2179x}$ ($R^2 = 0.93$).

Právě zastoupením dvou výše jmenovaných „plodných“ autorů v subkorpusu ES2 by tak bylo možné částečně vysvětlovat nižší celkové hodnoty oproti vysoce úspěšnému ES1. U německých subkorpusů pak můžeme usuzovat i na vliv jiného versifikačního systému (tónický verš), resp. odlišné metody rytmické analýzy (rytmické typy). U anglických subkorpusů se vliv počtu vzorků neprojevuje – v EN1 patří naopak nejobjemnější autorské dílo (George Gordon Byron; 46 vzorků) k lépe rozpoznatelným; v EN2 žádný z autorů počtem vzorků nevybočuje. Můžeme ale předpokládat, že tu do hry vstupují určité jazykově typologické rozdíly.

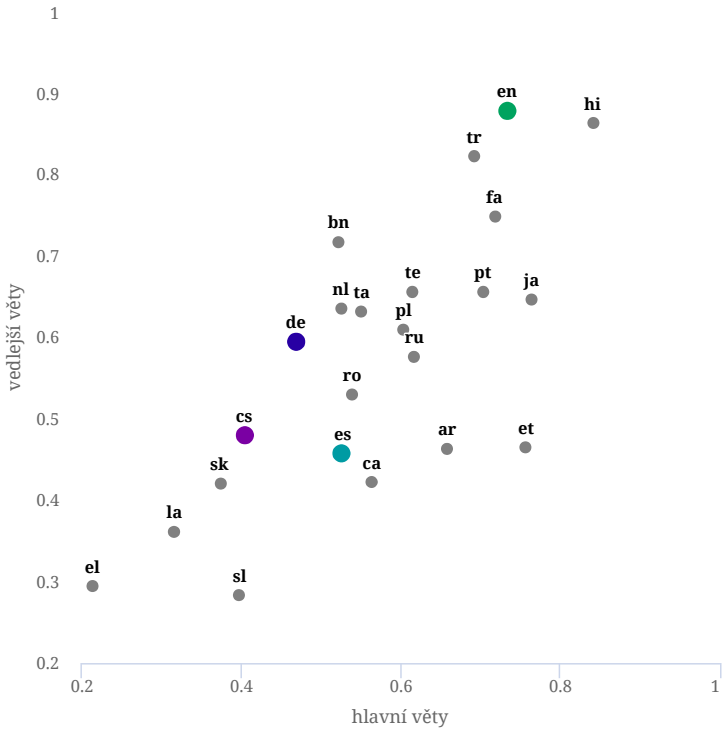
Vyjděme z předpokladu, že čím volnější má jazyk slovosled, tím více prostoru má autor pro rytmickou stylizaci svých veršů. Mezi pevností slovosledu a rozpoznatelností autorského stylu na základě rytmických rysů tak můžeme očekávat určitou míru asociace. Tradiční typologizace i empirická data (OBR. 3.4) přitom ukazují, že je to očekávání vcelku oprávněné.

Vedle toho můžeme předpokládat, že na výsledky může mít vliv i (s pevností slovosledu související) míra rozvinutosti flexe, u níž můžeme předpokládat vliv na bohatost rýmového slovníku. Jednoduše řečeno: čím víc gramatických sufixů se v jazyce objevuje, tím větší je repertoár rýmových párů, z nichž může básník vybírat.

Tuto hypotézu jsme – stejně jako v dřívější studii zabývající se rýmem v české, anglické a francouzské poezii (Plecháč 2018) – testovali na našich datech pomocí *type-token ratio* rýmových párů (r-TTR). Bohatost rýmového slovníku tak byla měřena jako podíl počtu různých rýmových párů (# r-TY) a celkového počtu rýmů (# r-TO):

$$\text{r-TTR} = \frac{\text{\# r-TY}}{\text{\# r-TO}} \qquad (3.1)$$

Protože metrika *type-token ratio* je obecně silně závislá na velikosti analyzovaného korpusu (srov. např. Popescu et al. 2009: 231–248), byly provedeny dvě sady experimentů s náhodným výběrem stejně velkých vzorků z každého ze čtyř korpusů (CS, DE, EN, ES). V prvním z nich bylo z každého korpusu vybráno 30 náhodných vzorků čítajících 10 000



OBR. 3.4: Modelování pevnosti slovosledu ve 23 jazycích provedené Kuboněm, Lopatkovou a Hercigem (2016). Eukleidovské vzdálenosti mezi vektory určenými četnostmi vět se strukturou SVO (subjekt–predikát–objekt), SOV, VOS, VSO, OVS a OSV a vektory určenými jejich rovnoměrně rozdělenými pravděpodobnostmi. Jazykové kódy dle ISO 639-1.

rýmových párů (výběr s opakováním) a u každého z nich spočtena hodnota r-TTR. V druhém případě byla procedura zopakována se vzorky čítajícími 50 000 rýmových párů.

Výsledky (OBR. 3.5) podporují původní hypotézu: vysoce flektivní čeština vykazuje u obou sad experimentů znatelně vyšší hodnoty r-TTR než angličtina s velmi omezenou flexí, zatímco středně flektivní němčina a španělština stojí mezi nimi. Ve všech případech zůstávají hodnoty naměřené ve vzorcích z jednoho jazyka poměrně stabilní.

Rozdíly mezi výsledky křížových validací napříč jednotlivými jazyky můžeme tedy přičítat spolupůsobení přinejmenším tří faktorů:

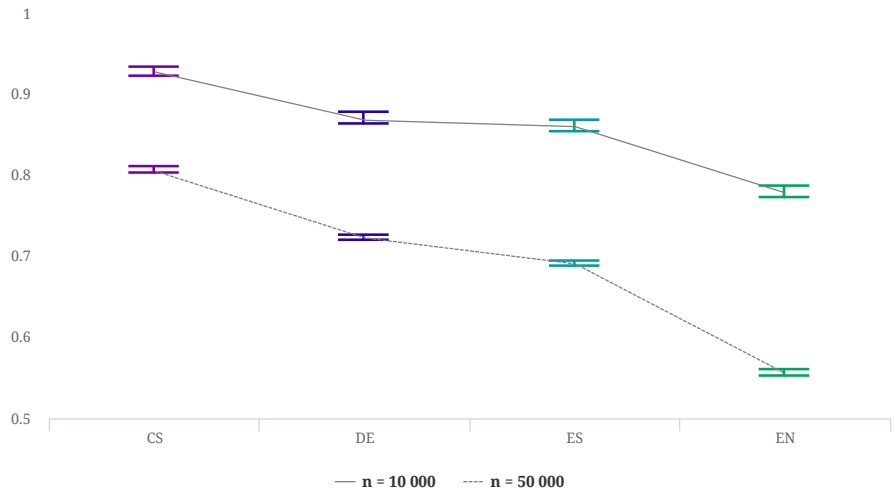
(1) počet vzorků od jednotlivých autorů, který často negativně koreluje s úspěšností;

(2) míra pevnosti slovosledu, která může omezovat rytmický repertoár;

(3) míra flektivnosti jazyka, která může ovlivňovat velikost rýmového repertoáru.

V úvahu samozřejmě přichází řada dalších, zejm. kulturně-historických faktorů, které se už ale ocitají mimo rámec této práce.[26]

26 Např. na základě obecně sdílených literárněhistorických charakteristik bychom mohli u romantických autorů předpokládat větší snahu rytmicky (či jakkoliv jinak) individualizovat svoji tvorbu než např. u autorů barokních.



OBR. 3.5: *Type-token ratio* rýmových párů; měření na 30 vzorcích čítajících 10 000 rýmových párů a 30 vzorcích čítajících 50 000 rýmových párů. Chybové úsečky zobrazují minima a maxima naměřených hodnot.

3.2.2 Klasifikační síla rysů

Kromě rozdílů mezi úspěšnostmi jednotlivých korpusů je vhodné se podívat i na to, jakou měrou přispívají ke klasifikaci jednotlivé rysy. Bez toho nelze vyloučit možnost, že některé z nich jsou pro atribuci zcela irelevantní; srov. např. i krajní variantu, že „ryze versologické“ rysy (rytmus a rým) nepřinášejí žádnou informaci a klasifikaci řídí pouze frekvence hlásek (v jazycích s fonologickým pravopisem jako čeština nebo španělština tedy rysy velmi blízké obvyklým stylometrickým ukazatelům – frekvencím grafémů).

Klasifikační sílu určitého rysu, jak bylo ukázáno v kap. 1.4.3, lze aproximovat souřadnicemi normálového vektoru v jemu odpovídající dimenzi. Abychom zjistili, jakou měrou přispívají jednotlivé rysy ke klasifikaci konkrétních autorů, zopakovali jsme výše popsanou sadu experimentů tak, že v každé z 30 iterací nebyla provedena křížová validace, ale veškerá data byla využita pro natrénování klasifikátorů pro jednotlivé autory strategií *one-vs.-rest* (v každé iteraci bylo tedy konstruováno 5 nadrovin).[27] Tímto způsobem jsme pro každého autora získali sadu až 30 normálových vektorů (v závislosti na tom, kolikrát se objevil v náhodném výběru).

Protože souřadnice normálových vektorů mohou nabývat jak kladné, tak záporné hodnoty, přičemž pro klasifikační sílu rysů není určující samotná hodnota, ale její vzdálenost od nuly, byly jednotlivé rysy napříč iteracemi ohodnoceny skóry spočtenými jako průměrná hodnota *druhých mocnin* jim odpovídajících souřadnic. Z nich bylo pak vybráno vždy 30 rysů s nejvyššími skóry (tj. těch které nejvíc přispívají k rozpoznání daného autora).

Vzhledem k tomu, že celkový počet rysů se pohybuje řádově ve stovkách, je vhodné se zaměřit na celé skupiny, tj. např. souhrnně „frekvence hlásek“ (nikoliv samostatné hodnoty pro jednotlivé hlásky), „rytmické bigramy“ (nikoliv samostatné hodnoty pro S_1W_1~00, S_1W_1~10…). TAB. 3.7 uvádí pro každého autora relativní zastoupení skupin rysů mezi 30 rysy s nejvyššími skóry.

27 Strategie *one-vs.-rest* byla v tomto případě zvolena proto, že (robustnější a častěji užívaná) strategie *one-vs.-one* konstruuje pro každého autora více nadrovin a interpretace souřadnic normálových vektorů je tak znatelně komplikovanější.

		(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)
		r-2-gram	**r-3-gram**	**r-4-gram**	**rým (1)**	**rým (2)**	**rým (3)**	**rým (4)**	**hlásky**
CS1	Čelakovský	0,03	0,07	0,1	0,03	0,2	**0,4**	0,13	0,03
	Havelka	0,1	0,1	**0,23**	0,03	0,17	0,17	0,13	0,07
	Hněvkovský	0,07	0,07	0,1		**0,27**	0,2	0,1	0,2
	Kulda	0,1	0,17	0,2	0,03	0,13	**0,23**		0,13
	Nejedlý	0,03	0,07	0,13		0,2	0,2	0,07	**0,3**
	Picek	0,17	0,2	**0,27**	0,03	0,13	0,13	0,03	0,03
	Pohan		0,2	**0,33**	0,03	0,2	0,2		0,03
	Tablic	0,07	0,13	**0,2**	0,13	0,17	0,1	0,17	0,03
	Vinařický	0,07	0,17	**0,23**	0,1	0,17	0,07	0,07	0,13
CS2	Čech	0,03	0,03		0,03	**0,4**	0,3	0,03	0,17
	Kvapil	0,07	0,13	0,2	0,07	**0,33**	0,1		0,1
	Mokrý	0,17	0,2	0,2	0,03	**0,23**	0,03	0,03	0,1
	Nečas		0,13	0,2	0,03	**0,47**	0,03	0,03	0,1
	Sládek	0,03	0,07	0,13	0,07	**0,43**	0,1		0,17
	Uden		0,07	0,17		**0,4**	0,17	0,03	0,17
	Vrchlický					**0,63**	0,3	0,03	0,03
CS3	Klášterský	0,13	0,07	0,17		**0,43**	0,17		0,03
	Kvapil	0,1	0,1	0,07	0,07	**0,33**	0,2		0,13
	Leubner	0,13	0,17	**0,23**		0,2	0,17		0,1
	Machar	0,13	0,1	0,1	0,1	**0,2**	0,13	0,1	0,13
	Sova	0,07	0,17	**0,33**		0,23	0,1	0,03	0,07
ES1	Acunya	0,03	0,03	0,1	0,03	0,07	**0,53**		0,2
	Borja	0,23	0,3	**0,37**			0,07		0,03
	Cetina	0,1	0,2	**0,3**	0,03		0,27		0,1
	Góngora	0,03	0,07	0,17	0,17	0,07	**0,27**		0,23
	Herrera	0,03	0,1	**0,23**	0,03	0,17	**0,23**		0,2
ES2	Argensola			0,07	0,13	0,03	**0,57**		0,2
	Quevedo	0,07	0,03	0,07	0,23		**0,57**		0,03
	Rojas	0,07	0,13	0,27	0,03	0,07	**0,3**		0,13
	Tassis y P.	0,07	0,07	0,1	0,1	0,03	**0,5**		0,13
	Ulloa y P.		0,1	**0,27**	0,2		0,23		0,2
	Vega	0,03	0,03	0,03		0,07	**0,77**		0,07
EN1	Byron		0,03	0,2	0,07	0,3	**0,37**	0,03	
	Coleridge			0,03	0,03	0,17	**0,5**	0,03	0,23
	Keats			0,03		0,5	**0,33**		0,13
	Shelley	0,03	0,07	0,23		0,2	**0,33**		0,13
	Wordsworth		0,03	0,03		0,47	**0,4**	0,03	0,03
EN1	Bierce	**0,2**	**0,2**	**0,2**	0,07	0,1	0,1		0,13
	Browning	0,07	0,1	0,2	0,03	0,2	**0,37**		0,03
	Hardy	0,03	0,13	0,17	0,03	0,07	**0,3**	0,03	0,23
	Lowell	0,03	0,07	0,03	0,03	0,37	**0,4**	0,03	0,03
	Wilde	0,03	0,1	0,13	0,13	**0,2**	**0,2**	0,07	0,13
				r-typy	**rým (1)**	**rým (2)**	**rým (3)**	**rým (4)**	**hlásky**
DE1	Brockes			0,23	0,07	0,23	**0,43**		0,03
	Drollinger			0,1	0,03	0,3	**0,37**	0,03	0,17
	Gottsched			0,07	0,07	0,23	**0,5**		0,13
	Kuhlmann			**0,3**	0,03	0,27	0,2		0,2
	Neukirch			0,37	0,03	0,03	**0,5**		0,07
	Tersteegen			0,13		0,13	**0,5**	0,03	0,2
DE2	Goethe			0,23	0,03	0,07	**0,5**	0,03	0,13
	Jacobi			**0,27**		**0,27**	**0,27**	0,03	0,17
	Müller			0,3	0,03	0,17	**0,37**		0,13
	Pfeffel			0,2		0,17	**0,33**	0,17	0,13
	Wieland			**0,53**	0,07	0,03	0,33		0,03
DE3	Bernhardi			0,3		0,27	**0,33**	0,03	0,07
	Eichendorff			0,23	0,03	0,13	**0,53**	0,03	0,03
	Grillparzer			**0,3**	0,07	0,2	**0,3**		0,13
	Müller			**0,33**	0,03	0,2	**0,33**	0,03	0,07
	Schenkendorf			**0,47**	0,17	0,13	0,17		0,07
	Schulze			**0,37**	0,03	0,13	0,33		0,13
	Tieck			**0,33**		0,07	**0,33**	0,03	0,23

TAB. 3.7: Klasifikační síla skupin rysů ((I–III) rytmické *n*-gramy / rytmické typy; (IV) slabičné délky rýmových slov; (V) hláskové složení rýmů; (VI) morfologické charakteristiky rýmů; (VII) přízvukové charakteristiky rýmů; (VIII) frekvence jednotlivých hlásek). Tabulka udává relativní četnosti rysů z jednotlivých skupin mezi 30 rysy s nejvyšším skóre (souřadnice normálového vektoru; srov. kap. 1.4.3). Nejvyšší hodnota je tučně zvýrazněna.

U českých subkorpusů dosahuje nejvyšších hodnot hláskové složení rýmových slov (V) – u Vrchlického tvoří dokonce více než 60 % ze sledovaných rysů, u ostatních jazyků se pak u rýmu ukazují jako nejproduktivnější morfologické charakteristiky (VI). U všech jazyků hrají důležitou roli rytmické charakteristiky (I–III), v případě rytmických *n*-gramů (CS, ES, EN) pak především rytmické tetragramy. Relativně nízké hodnoty napříč subkorpusy vykazují slabičné délky rýmových slov (IV) a přízvukové charakteristiky rýmů (VII).

Za pozornost stojí, že u subkorpusů ES1 a ES2 vykazuje druhá zmíněná skupina (VII) pouze nulové hodnoty (ve skutečnosti jsou zde dokonce všechny souřadnice rovny nule). Vysvětlení je jednoduché: konstantním rysem španělského hendekasylabu je zakončení sekvencí přízvučná slabika – nepřízvučná slabika, např.:

¡Peñascos Altos, de la mar batidos,
ryt. realizace: 0 1 0 1 0 0 0 1 0 1 0
de nubes coronadas las cabezas,
ryt. realizace: 0 1 0 0 0 1 0 0 0 1 0
donde se rompen en diversas piezas
ryt. realizace: 0 0 0 1 0 0 0 1 0 1 0
cristales espumosos resistidos
ryt. realizace: 0 1 0 0 0 1 0 0 0 1 0
(Lope de Vega)

Z tohoto pravidla neexistuje v korpusu ES jediná výjimka. Nulová variabilita rytmických charakteristik rýmových slov má tak přirozeně za důsledek nulovou využitelnost při klasifikaci. Žádnou ze skupin ale nelze označit za celkově dominantní a žádná skupina se neprojevuje jako celkově irelevantní.

3.3 Srovnání s obvyklými stylometrickými modely

Cílem druhé sady experimentů bylo jednak testovat úspěšnost versologických modelů oproti modelům založeným na obvyklých stylometrických rysech (dále lexikálně-morfologické modely), jednak testovat úspěšnost modelů kombinujících rysy z obou zmíněných skupin.

3.3.1 Výběr rysů a počet analyzovaných jednotek

V prvním kroku jsme se pokusili vybrat nejvhodnější lexikálně-morfologické modely pro naše data. Vzhledem k tomu, že v tomto případě nebylo třeba omezovat materiál na konkrétní veršový rozměr, bylo možné pracovat s objemnějšími subkorpusy než u versologických modelů (TAB. 3.8).

Testování bylo provedeno analogicky k versologickým modelům: v 30 iteracích bylo z každého subkorpusu náhodně vybráno 5 autorů, u každého z nich pak náhodně vybráno 10 vzorků. V každé iteraci byla provdena *leave-one-out* křížová validace SVM modelů založených na následujících rysech:

(1) četnosti *n* nejfrekventovanějších slovních tvarů,

(2) četnosti *n* nejfrekventovanějších lemmat,

(3) četnosti *n* nejfrekventovanějších znakových bigramů,

(4) četnosti *n* nejfrekventovanějších znakových trigramů,

(5) četnosti *n* nejfrekventovanějších znakových tetragramů,

zkratka	narozeni	# autorů	# vzorků
CS1'	1760–1820	41	1295
CS2'	1840–1855	31	2613
CS3'	1860–1870	32	2341
DE1'	1650–1699	14	776
DE2'	1730–1754	15	695
DE3'	1760–1794	23	1401
ES1'	1500–1560	6	122
ES2'	1561–1599	6	304
EN1'	1750–1799	6	434
EN2'	1800–1869	20	304

TAB. 3.8: Objemy rozšířených subkorpusů.

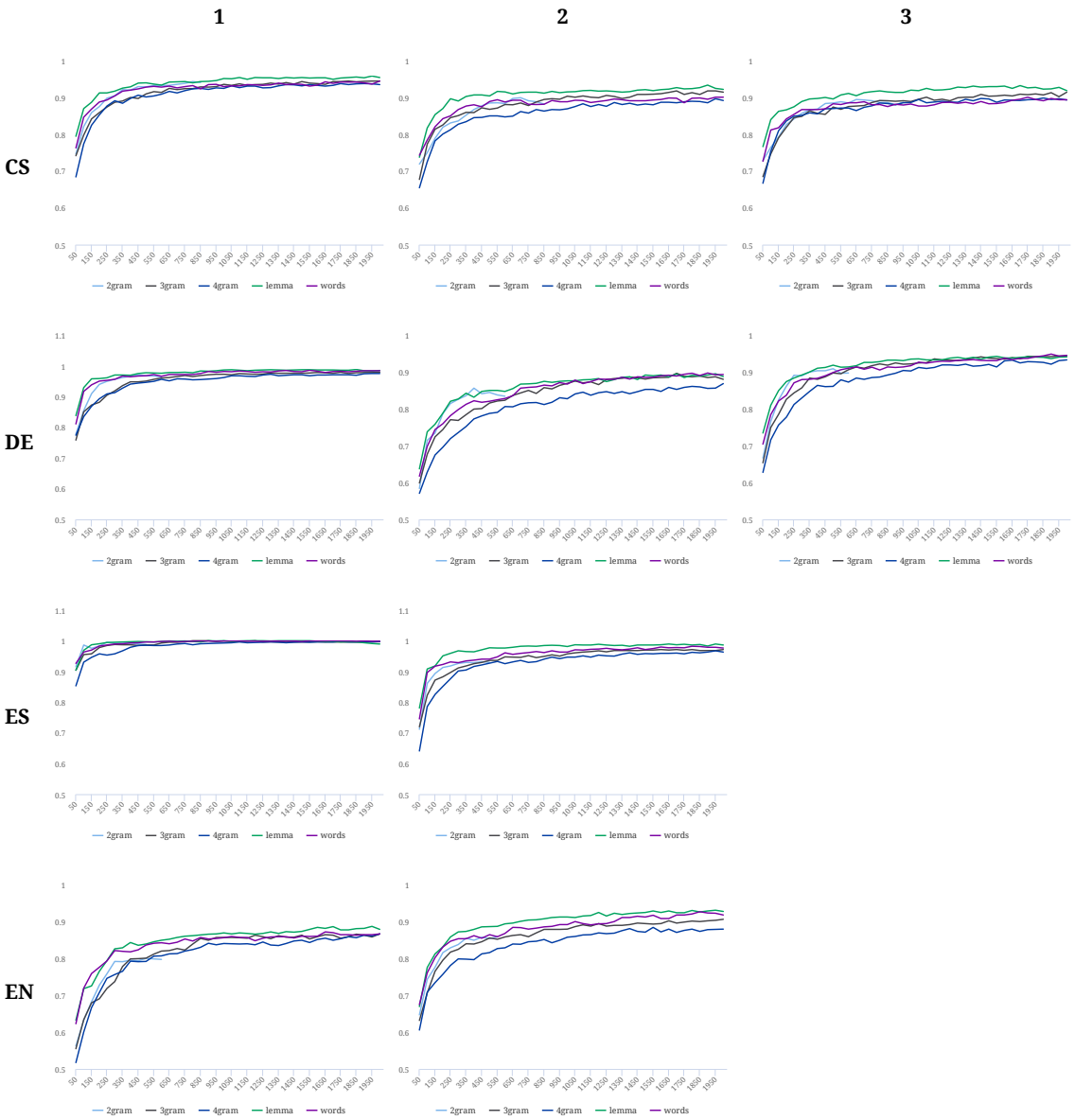
a to vždy pro čtyřicet různých nastavení $n \in \{50, 100, 150, \ldots, 2000\}$. Celkem tak bylo provedeno $30 \times 40 \times 10 = 12\,000$ křížových validací.

Výsledky (OBR. 3.6) potvrdily obecně známý jev: vztah mezi počtem analyzovaných jednotek (n) a úspěšností atribuce připomíná logaritmickou závislost; zpočátku strmě narůstá a po dosažení určité hodnoty se stabilizuje (srov. Eder 2011; Rybicki-Eder 2011; Smith-Aldridge 2011). Tato hodnota se přitom zdá být podobná pro všechny analyzované rysy v rámci jednoho subkorpusu, napříč subkorpusy se ale značně odlišuje (zatímco u ES1' úspěšnost kulminuje už při $n \approx 250$, např. u CS3' nebo DE2' úspěšnost narůstá až k nejvyšším hodnotám n).

Dále výsledky ukázaly, že ve všech sledovaných korpusech vykazují lemmata vyšší úspěšnost než slovní tvary nebo znakové n-gramy. (Hypotéza, že ve flektivních jazycích jsou vhodnějším ukazatelem znakové n-gramy (srov. kap. 1.3.2) se tak nepotvrdila.) Ze skupiny znakových n-gramů pak obvykle dosahují nejvyšší úspěšnost trigramy.

Na první pohled by se tak mohlo zdát, že nejspolehlivější metodu představuje analýza nejfrekventovanějších lemmat při co možná nejvyšší hodnotě n. Je ale třeba vzít v potaz, že vyšší hodnoty n s sebou nesou i zvyšující se riziko overfittingu: čím méně frekventované jednotky bereme v potaz, tím větší je riziko, že klasifikátor ve skutečnosti nerozpoznává obecné charakteristiky autorského stylu, ale dílčí témata. Typickým příkladem je jeden z experimentů provedený se vzorky dvou autorů z korpusu CS3', Sigismunda Boušky a Františka Cajthamla-Liberté. Podíváme-li se v tomto případě na 10 rysů s největší klasifikační silou (TAB. 3.9), zjistíme, že klasifikátor rozpoznává především tematické odlišnosti (katolická × proletářská poezie) a že by při aplikaci na případné jinak tematicky zaměřené básně těchto autorů patrně nebyl příliš užitečný.

Tento předpoklad jsme se pokusili ověřit dalším experimentem. Cílem bylo zjistit úspěšnost modelů natrénovaných na epických básních při klasifikaci básní lyrických a naopak (přičemž předpokládáme, že literární žánr má v tomto ohledu podobný efekt jako tematika). Pro tyto potřeby bylo vybráno 5 autorů z korpusu CS2' a 5 autorů z korpusu EN1', z jejichž děl lze dobře vyčlenit dostatek lyrických a epických básní. Konkrétně se jednalo o české autory spojované s tzv. ruchovsko-lumírovskou školou: Svatopluk Čech, Eliška Krásnohorská, Rudolf Pokorný, Ladislav Quis, Jaroslav Vrchlický a anglické autory období romantismu: George Gordon Byron, John Keats, Percy Bysshe Shelley, Robert Southey, William Wordsworth (výběr a rozdělení textů na lyrické a epické je uvedeno v TAB. 3.10).



OBR. 3.6: Výsledky křížových validací modelů založených na 50, 100, 150, ..., 2000 nejfrekventovanějších znakových bigramů, znakových trigramů, znakových tetragramů, lemmat a slovních tvarů.

	Bouška	Cajthaml-Liberté
1	svatý	práce
2	boží	černý
3	nebesa	bída
4	jaký	lid
5	Kristus	chléb
6	mluvit	zítra
7	volat	dělník
8	nebeský	ležet
9	otec	ruch
10	klín	již

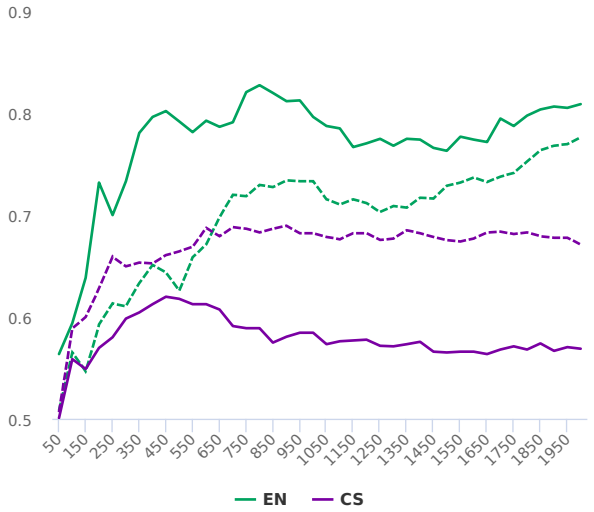
TAB. 3.9: Lemmata, jejichž frekvence vykazují nejvyšší klasifikační sílu pro Sigismunda Boušku a Františka Cajthamla-Liberté.

	Autor (# vzorků lyriky / # vzorků epiky)	Lyrika	Epika
CS2'	Čech (23/20)	*Jitřní písně; Nové písně*	*Václav z Michalovic; Lešetínský kovář; Petrklíče*
	Krásnohorská (37/25)	*Vlny v proudu; Letorosty*	*Vlaštovičky; Šumavský Robinson; Zvěsti a báje*
	Pokorný (17/9)	*S procitlým jarem; Vlasti a svobodě*	*Mrtvá země*
	Quis (11/10)	*Písničky*	*Hloupý Honza; Třešně*
	Vrchlický (42/34)	*Dni a noci; Hořká jádra; È morta*	*Hilarion; Sfinx; Poutí k Eldoradu*
EN1'	Byron (13/28)	*Epitaph on a Beloved Friend; Adrian's Address to the Soul when Dying; A Fragment; To Caroline; On a Distant View of the Village and School of Harrow on the Hill; Thoughts Suggested by a College Examination; To Mary; On the Death of Mr. Fox; To a Lady who Presented...; To a Beautiful Quaker; To Lesbia!; To Woman; An Occasional Prologue; To Eliza; The Tear; Reply to Some Verses...; Granta. A Medley; To the Sighing Strephon; The Cornelian; To M----; Lines Addressed to a Young Lady; To the Eyes of Miss A----H----; To a Vain Lady; To Anne; Egotism; To the Author of a Sonnet; On Finding a Fan; Farewell to the Muse; To an Oak at Newsted; On Revisiting Harrow; To my Son; Queries to Casuists; Song; There was a Time, I need not a Name; And wilt thou Weep when I am Low?; Epistle to a Young Nobleman in Love; Remid me not, Remind me not; To a Youthful Friend; Lines Inscribed upon a Cup...; Well! Thou Art Happy; Inscription on the Monument...; To a Lady; Fill the Goblet Again; Stanzas to a Lady, on Leaving England*	*The Prisoner of Chillon; Lara; Corsair*
	Keats (9/9)	*Sonnets I–XVIII; To Some Ladies; On Receiving a Curious Shell; To****; To Hope; To George Felton Mathew; To my Brother George; To Charles Cowden Clarke; Sleep and Poetry; Ode to a Nightingale; Ode to Grecian Urne; Ode to Psyche; Fancy; Ode to Melancholy*	*Hyperion; Endymion; Lamia*
	Shelley (9/13)	*Ode to the West Wind; The Sensitive Plant; The Cloud; To a Skylark; Ode to the Liberty; Hymn of Apollo; The Question; Ode to the Naples; Autumn: A Dirge; The Waning Moon; To the Moon; Death; Liberty; Summer and Winter; The Tower of Famine; An Allegory; The World's Wanderers; Sonnet; Ozymandias; To the Nile; To Harriet; To Mary Wollstonecraft Godwin; To----; Mutability; On Death; A Summer Evening Churchyard*	*The Revolt of Islam*
	Southey (9/10)	*Sonnet I–X; Sappho; Ode; Written on Sunday Morning; On the Death of a Favourite Old Spaniel; To Horror; The Soldier's Wife; The Widow; To the Chapel Bell; The Race of Banquo; Musings on a Landscape; Mary; Donica; Rudiger; Hymn to the Penates*	*The Vision of the Maid of Orleans; Ballads*
	Wordsworth (10/12)	*Incident Characteristic of a Favourite Dog; Tribute to the Memory of the Same Dog; To the Daisy; Elegiac Stanzas; Elegiac Verses; When to the Attractions of the Busy World; Alas! What Boots...; On the Final Submission...; The Martial Courage...; And is it Among...; O'er the Wide Earth...; Hail!, Zaragoza.!...; Say, what is Honour?...; Brave Schill!...; Call not the Royal Swede Unfortunate; Look now on that Adventurer...; Is there a Power...; Weep not, Beloved Friends!...; Perhaps Some Needful Service...; O thou who Movest...; There Never Breathed a Man...; True is it...; Destined to War...; O Flower of All...; Not Without Heavy Grief...; Pause, Curteous Spirit...; Ah! Where is Palafox?...; In Due Observance...; Feeling of a Noble Biscayan...; On a Celebrated Event...; Upon the Same Event; The Oak of Guernica; Indignation of a High-Minded Spaniard; Avaunt All Specious Plinancy of Mind; O'eweeing Statesmen...; The French and the Spanish Guerillas; Maternal Grief; Characteristics of a Child; The Power of Armies...; Here Pause...; Epistle to Sir George Howland...; Upon Perusing the Foregoing Epistle...; Upon the Sight of a Beautiful Picture; To the Poet, John Dyer; Song for the Spinning Wheel; Composed on the Eve...; Water-Fowl; View from the Top of Black Comb; Written with a Slate Pencil...; November*	*Peter Bell; The Waggoner; The White Doe of Rylstone*

TAB. 3.10: Vzorky lyriky a epiky z korpusů CS2 a EN1. V prvním případě jsou uvedeny celé básnické sbírky, v druhém případě konkrétní básně, resp. epické celky.

Analogicky k výše popsanému experimentu bylo v 30 iteracích u obou pětic autorů na náhodně vybraných vzorcích epiky (9 vzorků pro každého autora)[28] natrénováno 40 modelů (pro 50, 100, 150, …, 2000 nejfrekventovanějších lemmat) a jejich úspěšnost testována na náhodně vybraných vzorcích lyriky (9 vzorků pro každého autora). Celý proces byl pak zopakován s trénováním na lyrických vzorcích a testováním na vzorcích epických.

Výsledky (OBR. 3.7) ukazují značně odlišný trend oproti předchozímu experimentu. U českých autorů dosahuje úspěšnost rozpoznávání lyrických vzorků maxima při n = 450 a následně klesá, úspěšnost rozpoznávání epických vzorků dosahuje maxima při n = 600 a následně zůstává stabilní. U anglických autorů dosahuje úspěšnost rozpoznávání lyrických vzorků maxima při n = 800, následně klesá až k hodnotě n = 1500, odkud opět lehce narůstá, a pouze u rozpoznávání epických vzorků můžeme pozorovat více méně konstantní nárůst úspěšnosti.[29]



OBR. 3.7: Úspěšnost rozpoznávání vzorků lyriky při trénování modelů na vzorcích epiky (plná čára) a rozpoznávání vzorků epiky při trénování modelů na vzorcích lyriky (přerušovaná čára) za použití 50, 100, 150, …, 2000 nejfrekventovanějších lemmat.

Jako optimální reference se tak pro versologické modely zdají být modely založené na 500 nejfrekventovanějších lemmatech – na hladině, kdy už úspěšnost ve všech subkorpusech dosahuje maxima nebo následně narůstá jen mírně, a zároveň lze ještě odhadovat, že riziko overfittingu je poměrně nízké. *Nota bene*, že 500rozměrné vektory patří ve stylometrii k často užívaným právě i v oblasti básnických textů (např. Craig–Kinney 2009; Smith-Aldridge 2011). Pro srovnání se dále zaměříme i na dvě nižší hladiny, s nimiž se setkáme (mimo jiné) ve dvou klasických studiích: n = 150 (Burrows 2002) a n = 250 (Koppel–Schler 2004).

28 Tj. počet vzorků na jednoho autora, který zbývá v trénovacích datech při křížové validaci po náhodném vyřazení. Srov. pozn. 23.

29 Rozdíl mezi klesající úspěšností u rozpoznávání lyriky a konstantní nebo narůstající úspěšností u rozpoznávání epiky lze pravděpodobně přičítat zejm. propriím. Při trénování na vzorcích epiky se u vyšších hladin n mezi analyzovaná lemmata často dostávají jména postav (při n = 2000 tak např. mezi rysy s nejvyšší klasifikační silou pronikají *Lara, Conrad* (Byron), *Endymion* (Keats), *Honza, Honzík* (Quis), *Hilarion* (Vrchlický), nebo *Václav* (Čech).

3.3.2 Výsledky

Při srovnávání versologických modelů a modelů založených na četnostech nejfrekventovanějších lemmat (nadále „lexikální modely“) byl uplatněn stejný postup jako při testování versologických charakteristik samotných: každý ze subkorpusů CS1–3, DE1–3, ES1–2 a EN1–2 byl v 30 iteracích zredukován na 50 vzorků (náhodný výběr 5 autorů a 10 vzorků pro každého z nich), přičemž byla provedena křížová validace:

(1) versologických modelů (tytéž rysy jako v kap. 3.2);

(2) lexikálních modelů (n = 500);

(3) modelů založených na kombinaci obou výše zmíněných sad rysů (spojení obou vektorových prostorů).

Celý proces byl pak zopakován s použitím lexikálních modelů při n = 150 a n = 250.

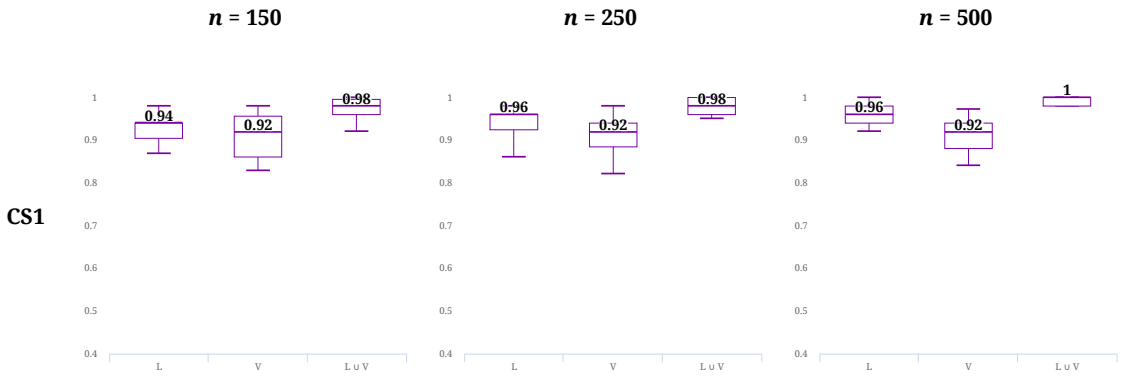
Výsledky (OBR. 3.8) ukazují, že:

(1) u lexikálních modelů narůstá dle očekávání (kap. 3.3.1) úspěšnost u každého subkorpusu směrem od nižších hodnot n k vyšším;

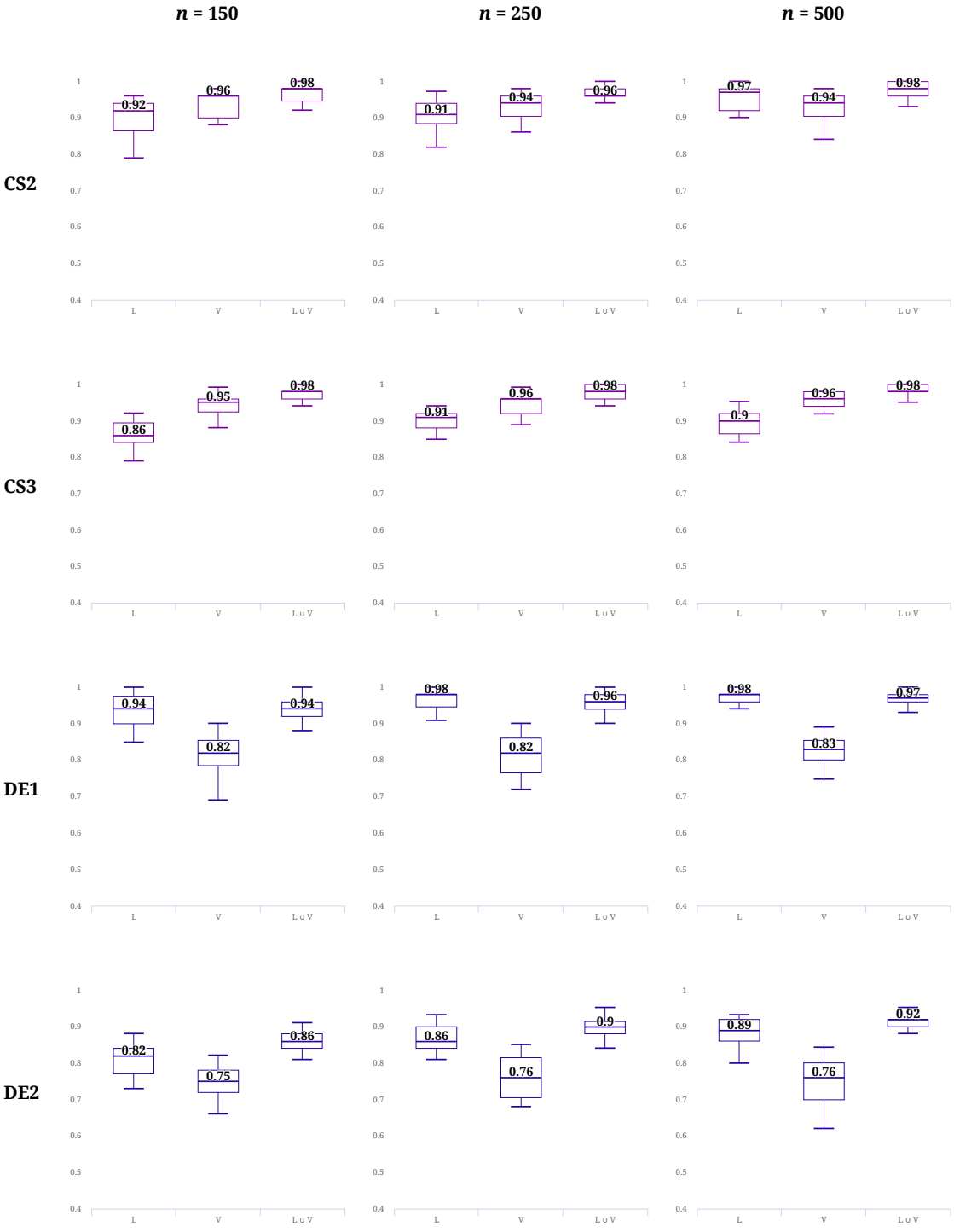
(2) versologické modely vykazují napříč náhodnými výběry více méně konstantní úspěšnost;

(3) v pěti případech (CS2 při $n \in \{150, 250\}$ a CS3 při $n \in \{150, 250, 500\}$) vykazují versologické modely vyšší úspěšnost než modely lexikální, v ostatních případech je úspěšnost nižší (nejvyšší propad vykazují anglické subkorpusy);

(4) kombinované modely založené na spojení obou vektorových prostorů vykazují u všech českých subkorpusů a subkorpusů DE2 a DE3 na všech třech hladinách n vyšší úspěšnost než modely založené pouze na jednom z nich (ve všech případech jsou rozdíly mezi kombinovanými modely a lexikálními modely statisticky významné na konvenční hladině významnosti α = 0,05; srov. TAB. 3.11). U subkorpusů DE1, ES1 a ES2 nepřinášejí kombinované modely oproti lexikálním modelům žádný nárůst úspěšnosti (s výjimkou ES1 při n = 150), u subkorpusů EN1 a EN2 dochází naopak k významnému poklesu úspěšnosti.



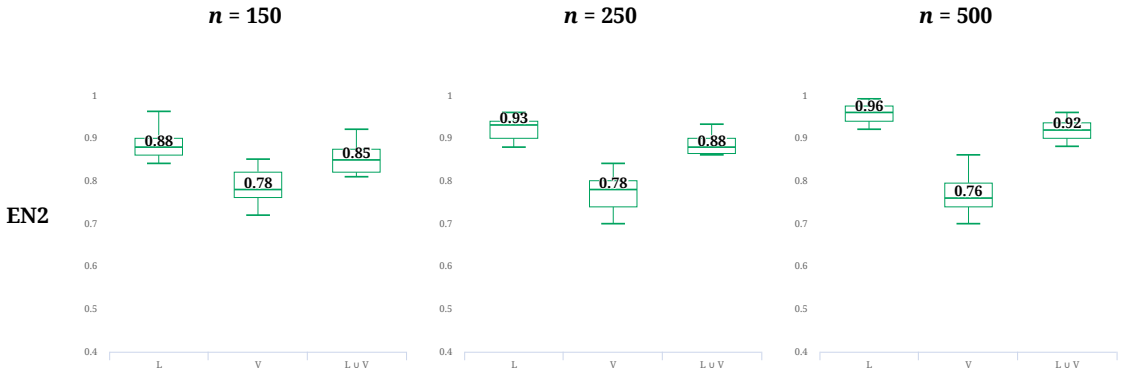
OBR. 3.8...


n = 150
n = 250
n = 500
CS2
CS3
DE1
DE2
L
V
L ∪ V
0.92
0.96
0.98
0.91
0.94
0.96
0.97
0.94
0.98
0.86
0.95
0.98
0.91
0.96
0.98
0.9
0.96
0.98
0.94
0.82
0.94
0.98
0.82
0.96
0.98
0.83
0.97
0.82
0.75
0.86
0.86
0.76
0.9
0.89
0.76
0.92

OBR. 3.8...


n = 150
n = 250
n = 500
DE3
ES1
ES2
EN1
L
V
L ∪ V
0.88
0.79
0.9
0.92
0.82
0.93
0.96
0.82
0.97
0.98
0.94
1
1
0.94
1
1
0.96
1
0.94
0.8
0.94
0.96
0.84
0.96
0.98
0.79
0.98
0.8
0.58
0.73
0.89
0.58
0.82
0.89
0.58
0.82

OBR. 3.8...



OBR. 3.8: Výsledky křížových validací modelů založených na četnostech 150, 250 a 500 nejfrekventovanějších lemmat (L), versologických rysech (V) a spojení obou vektorových prostorů (L ∪ V); 30 iterací s náhodným výběrem vzorků.

n	CS1	CS2	CS3	DE1	DE2	DE3	EN1	EN2	ES1	ES2
150	**< 10^{-4}**	**< 10^{-5}**	**< 10^{-5}**	0,3878	**< 10^{-4}**	**0,0013**	**< 10^{-4}**	**0,0030**	**< 10^{-4}**	0,6029
250	**< 10^{-4}**	**< 10^{-5}**	**< 10^{-4}**	0,1739	**0,0132**	**0,0347**	**< 10^{-4}**	**< 10^{-5}**	0,7963	0,3879
500	**< 10^{-4}**	**0,0042**	**< 10^{-5}**	0,3608	**0,0002**	**0,0077**	**< 10^{-5}**	**< 10^{-5}**	1	0,4638

TAB. 3.11: *P*-value rozdílů mezi úspěšnostmi modelů založených na četnostech nejfrekventovanějších lemmat a úspěšnostmi kombinovaných modelů (Wilcoxonův znaménkový test). Statisticky významné rozdíly (α = 0,05) jsou zvýrazněny tučně.

Kromě kombinace obou modelů stojí za pozornost i situace, kdy se lexikální model s versologickým modelem navzájem korigují, tedy přístup kdy při validaci mohou být výstupem klasifikace každého vzorku tři výsledky:

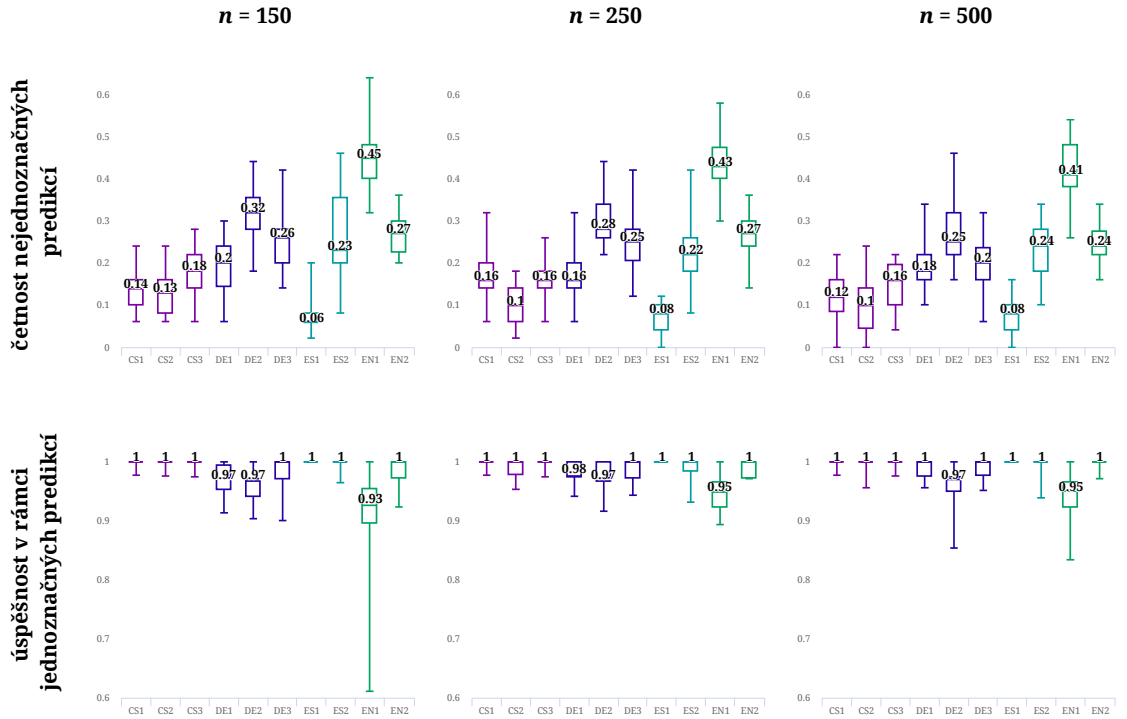
(1) správná predikce (oba modely predikují téhož správného autora);

(2) chybná predikce (oba modely predikují téhož nesprávného autora);

(3) nejednoznačná predikce (každý model predikuje jiného autora).

Výsledky testování tohoto přístupu (na vzorcích extrahovaných v 30 iteracích předchozího experimentu) jsou shrnuty na OBR. 3.9. Za cenu vyřazení některých vzorků coby nejednoznačných dochází ve všech případech k významnému navýšení úspěšnosti oproti lexikálním modelům, versologickým modelům i modelům kombinovaným (Wilcoxonův znaménkový test; α = 0,05). Výjimkou je ES1 při $n \in \{150, 250, 500\}$, kde u modelů založených na četnostech nejfrekventovanějších lemmat není co korigovat, a DE1 při n = 500.

Bylo by možné namítnout, že přístup, který zahazuje velkou část vzorků coby nejednoznačných (v případě EN1 někdy i více než polovinu), není příliš užitečný. To by ale platilo pouze v případě, kdy bychom oběma modelům přikládali stejnou váhu, a nikoli vezmeme-li lexikální model jako primární a versologický model pouze jako kontrolní, tzn. klasifikace je provedena prvním zmíněným a druhý pouze dodává další evidenci. Jinými slovy, pokud u konkrétního sporného textu predikuje lexikální model téhož autora jako model versologický, můžeme takovou atribuci obecně pokládat vždy za spolehlivější než situaci, kdy je daný autor predikován pouze lexikálním modelem.



OBR. 3.9: Četnosti nejednoznačných predikcí v rámci křížové validace (model založený na četnostech nejfrekventovanějších lemmat predikuje jiného autora než versologický model); zastoupení správných predikcí v rámci všech jednoznačných predikcí (oba modely predikují téhož autora); 30 iterací s náhodným výběrem vzorků.

3.4 Shrnutí

Výše prezentované výsledky ukazují, že versologické rysy představují smysluplný indikátor autorství:

(1) úspěšnost versologických modelů vysoce převyšuje hranici *random baseline;*

(2) úspěšnost versologických modelů ojediněle převyšuje i úspěšnost obvyklých lexikálních modelů;

(3) úspěšnost kombinovaných versologicko-lexikálních modelů je zpravidla vyšší než úspěšnost lexikálních modelů samotných;

(4) situace, kdy lexikální a versologický model predikuje stejného autora znamená téměř vždy vyšší spolehlivost atribuce než situace, kdy je predikce provedena pouze lexikálním modelem.

Zároveň se ukázalo, že úspěšnost versologických a kombinovaných modelů je značně závislá na konkrétní versifikaci. Jako nejspolehlivější se versologické modely ukázaly u českého sylabotónického verše a (v o něco menší míře) španělského sylabického verše. Poměrně dobrých výsledků dosáhly versologické modely i v případě německého tónického verše.

Nejhorší výsledky vykázaly versologické modely u anglického sylabotónického verše. To jsme se pokusili vysvětlit jednak pevností anglického slovosledu, který omezuje možnosti individualizace rytmu, jednak málo rozvinutou flexí, což má za důsledek chudší rýmový slovník oproti ostatním jazykům. I přesto ale mohou versologické rysy najít v anglické versifikaci uplatnění. Jak uvidíme později (kap. 4.1), přinejmenším v případě binární nebo ternární klasifikace a většího množství dat.

4 Aplikace

V poslední části práce budeme výše testované přístupy aplikovat na dva případy sporného autorství: (1) veršované drama *The Famous History of the Life of King Henry the Eight* – které jsme zmínili už v kap. 1.1 a 1.5 – poprvé otištěné v Prvním foliu her Williama Shakespeara, u nějž se ovšem předpokládá i autorská účast John Fletchera, případně i dalších autorů, a (2) básně publikované pod jménem Josefa Baráka (1833–1883), u nichž existuje hypotéza, že jejich skutečným autorem je Jan Neruda.

Tyto případy přitom představují zajímavý protipól jak z hlediska metodologie, tak z hlediska povahy dat. U hry *The Famous History of the Life of King Henry the Eight* se jedná o situaci, kterou jsme se až doteď zabývali: pro každou scénu máme konečnou množinu kandidátů a dostatek trénovacích dat (hry prokazatelně napsané Williamem Shakespearem, hry Johna Fletchera, případně Philipa Massingera), cílem je vybrat z této množiny nejpravděpodobnějšího autora. Jedná se tedy o *klasifikační úlohu*. U básní publikovaných pod jménem Josefa Baráka je situace odlišná: vzhledem k tomu, že žádné jiné prokazatelně Barákovy básně nemáme doloženy, je cílem ověřit, zda básně napsal, anebo nenapsal Jan Neruda. Jedná se tedy o *verifikační úlohu*.

Z hlediska dat se pak oba problémy vyznačují nutností spolehnout se jen na jeden typ versologických rysů (rytmické charakteristiky / charakteristiky rýmů) – na jedné straně stojí alžbětinské drama psané typickým blankversem (dostatek veršů v jednom rozměru, ale téměř úplná absence rýmů), na druhé straně sbírka rýmovaných básní s velmi pestrým metrickým repertoárem (dostatek rýmů, ale nedostatek veršů v jednom rozměru).

4.1 William Shakespeare a John Fletcher: *Henry VIII*

Veršované drama *The Famous History of the Life of King Henry the Eight* (nadále *H8*) publikované v prvním souboru her Williama Shakespeara (tzv. První folio; 1623) bylo až do poloviny 19. století považované za nesporné Shakespearovo dílo. Od té doby byla publikována řada hypotéz týkajících se jak dalších možných spoluautorů, tak Shakespearovo autorství zcela vylučujících. Následující výčet se omezuje na práce, které vycházejí z kvantitativní analýzy textu, příp. práce, na něž tyto kvantitativní analýzy navazují nebo se vůči nim polemicky vymezují.

4.1.1 Přehled dosavadních atribucí

Jako první vystoupil s hypotézou o smíšeném autorství *H8* tehdejší editor díla Francise Bacona James Spedding v článku *Who wrote Shakspere's Henry VIII.?* otištěném v časopise *Gentleman's Magazine and Historical Review* (1850). Spedding zde podává dlouhý výčet tematicko-motivických inkoherencí nalezených v textu hry a popisuje celkový čtenářský dojem z jednotlivých scén: „The result of my examination was a clear conviction that at

least two different hands had been employed in the composition of Henry VIII.; if not three; and that they had worked, not together, but alternately upon distinct portions of it“ (1850: 118). Spedding tak formuluje hypotézu rozdělující autorství mezi Williama Shakespeara a Johna Fletchera. Každou scénu přitom přisuzuje jednomu autorovi. Výjimkou je druhá scéna třetího jednání, kde první část přisuzuje Shakespearovi (po půlverš „What apetite you have“, tj. odchod krále ze scény), druhou Fletcherovi. Hlavním argumentem se stává nápadně vysoká variabilita poměru četností žensky zakončených veršů (např. „Till this time pomp was single, but now *married*“) a četností mužsky zakončených veršů (např. „The view of earthly glory: men might *say*“) napříč jednotlivými scénami (TAB. 4.1); Spedding uvádí, že nižší četnosti zhruba odpovídají hodnotám naměřeným ve dvou pozdních Shakespearových hrách (*Cymbeline*, *Winter's Tale*), zatímco vyšší četnosti zhruba odpovídají hodnotám naměřeným ve scéně ze čtvrtého jednání hry *Thierry and Theodoret* pravděpodobně napsané Fletcherem (vedle Fletchera se u této hry předpokládá spoluautorství Francise Beaumonta a Philipa Massingera).

Ještě tentýž rok reagoval na článek dopisem otištěným v *Notes and Queries* Samuel Hickson (1850). Uvádí, že nezávisle na Speddingovi dospěl k témuž rozdělení autorství mezi Shakespeara a Fletchera (včetně rozdělení druhé scény třetího jednání). Na rozdíl od Speddinga se zabývá i prologem a epilogem, přičemž oba přisuzuje Fletcherovi.

Jednání	Scéna	Podíl žensky zakončených veršů	Spedding-Hicksonova atribuce
	prolog	---	Fletcher
I	**1**	0,28	Shakespeare
	2	0,34	Shakespeare
	3, 4	0,58	Fletcher
II	**1**	0,59	Fletcher
	2	0,60	Fletcher
	3	0,38	Shakespeare
	4	0,31	Shakespeare
III	**1**	0,72	Fletcher
	2a	0,32	Shakespeare
	2b	0,59	Fletcher
IV	**1**	0,49	Fletcher
	2	0,59	Fletcher
V	**1**	0,39	Shakespeare
	2, 3[30]	0,53	Fletcher
	4	---	Fletcher
	5	0,60	Fletcher
	epilog	---	Fletcher

TAB. 4.1: Podíl žensky zakončených veršů v jednotlivých scénách *H8* dle Jamese Speddinga (1850: 121). III.2a označuje první část scény (po odchod krále ze scény), III.2b zbytek. Hodnoty pro V.4 Spedding neuvádí – je převážně tvořena prózou. Speddingem uváděné poměry jsou přepočítány na relativní četnosti. Autorství jednotlivých scén podle Speddinga (1850), resp. Hicksona (1850).

O čtvrtstoletí později se k Spedding-Hicksonově atribuci vrátili členové *New Shakspere Society* a v prvním svazku svých *Transactions* (1874) publikovali dva články přinášejících další doklady na její podporu:

30 Moderní edice nahrazují původní členění pátého jednání na čtyři scény členěním na pět scén: původní V.2 je rozdělena na V.2 a V.3 (počínaje veršem „Speak to the business, master-secretary“), z původního V.3 se tak stává V.4, z původního V.4 se stává V.5. V této práci se přidržujeme moderního členění.

1. John Kells Ingram zavádí jemnější rozlišení ženských zakončení na „light endings“ a „weak endings“ (viz pozn. 4) a udává poměr jejich četností kromě *H8* také pro všechny prokazatelně Shakespearovy hry a hru *Custom of the Country,* kterou označuje za bezpečně Fletcherovu.[31] Konstatuje, že pozdní Shakespearovo dílo se vyznačuje vysokou frekvencí typu „weak ending“, zatímco hra *Custom of the Country* naopak vysokou frekvencí typu „light ending“, a že poměry v *H8* odpovídají Spedding-Hicksonově atribuci: „An examination of the weak endings in *Henry VIII.* strikingly confirms the conclusions of Mr Spedding respecting the two different systems of verse which co-exist in that play. In the Shaksperian portion, as marked off by him, there are 45 light endings against 7 in Fletcher's part, and 37 weak endings against 1 in Fletcher's part“ (Ingram 1874: 453).
2. Frederick James Furnivall (1874b) dokládá, že každá scéna přisuzovaná Spedding-Hicksonovou atribucí Shakespearovi obsahuje víc případů neshody veršového a syntaktického členění než každá ze scén přisuzovaných Fletcherovi („stopt-line test“).

V roce 1885 přednesl Robert Boyle členům *New Shakspere Society* novou teorii, podle níž původní text Shakespearovy hry věnovaný králi Jindřichu VIII. shořel v roce 1613 při požáru divadla Globe[32] a hra otištěná v Prvním foliu je ve skutečnosti jejím přepracováním z pera Johna Fletchera a Philipa Massingera. Boyle postupuje téměř verš po verši a poukazuje na podobná slovní spojení a obrazy v různých prokazatelně Fletcherových a Massingerových dílech. Ve výsledku připisuje větší část hry (co do počtu veršů) Massingerovi, přičemž v řadě případů se části připisované jednomu autorovi nekryjí s hranicemi scén (viz OBR. 4.1, kde jsou znázorněny i všechny následující atribuce).

Na Boylovu teorii navázal Frederick Gard Fleay (1885, 1886), někdejší horlivý zastánce Spedding-Hicksonovy atribuce (Fleay 1874c) a toho času už bývalý člen *New Shakspere Society.* Fleay přijímá Fletcherovu a Massingerovu účast, ale některé z údajně Massingerových scén připisuje zpět Shakespearovi. Argumentuje přitom jednak stejně jako Boyle obrazností („The scenes between Henry and Katherine are beyond Massinger‘s powers […] Again, the one scene in which Anne Boleyn discloses her essentially mean nature is too subtle for Massinger“; Fleay 1885: 355), jednak přítomností alexandrinů v daných scénách („I have given a list of alexandrines of peculiar formation in i. 2, ii. 3, ii. 4 and I challenge Mr. Boyle to parallel these in the same number of lines in any part of Massinger‘s work. On the other hand, they are paralleled in »The Tempest,« »Winter‘s Tale,« &c.“; Fleay 1885: 355). O pár let později připisuje *H8* Shakespearovi, Fletcherovi i Massingerovi zároveň (s odlišným rozvržením a bez jakékoliv argumentace) i Ernest Henry Oliphant (1891: 326 –327).

Proti Massingerově účasti a na podporu Spedding-Hicksonovy atribuce vystoupil Ashley Horace Thorndike (1901: 39–44). Svoji argumentaci založil na poměru četností variant *them* a *‘em* v jednotlivých scénách: zatímco pro Shakespeara je typická spíš preference *them* (ve *Winter’s Tale* nachází Thorndike 18 % zkrácených tvarů), pro Fletchera je typická preference *‘em* (93–94 % zkrácených tvarů v *Bonduce* i *Woman’s Prize*). V částech připisovaných Spedding-Hicksonovou atribucí Shakespearovi přitom Thorndike nachází

31 Dnes se ovšem předpokládá spíš spoluautorství Johna Fletchera a Philipa Massingera (srov. Hoy 1957).

32 Z dochovaných pramenů vyplývá, že požár vypukl během inscenace hry s titulem nebo podtitulem *All is True* tematizující vládu Jindřicha VIII (srov. např. Oliphant 1927: 302).

23 % zkrácených tvarů, zatímco v částech připisovaných Fletcherovi 93 %. Tím vyvrací i možnost Massingerovy účasti: části, které Boyle, Fleay i Oliphant označují za Massingerovy, obsahují vždy několik případů zkrácených tvarů, kterým se Massinger důsledně vyhýbal (v sedmi jeho hrách nalezl Thorndike 210 výskytů *them* a žádný výskyt *'em*). Tato argumentace byla ale lichá. Jak Thorndike sám upozornil v erratech, později zjistil, že pracoval s vydáním Massingerových her, v němž editoři převedli všechny zkrácené tvary na nezkrácené (srov. Vickers 2004: 345).

Thorndikova práce nicméně podnítila Willarda Edwarda Farnhama (1916) k detailnějšímu výzkumu užívání kontrakcí. Farnham vyčleňuje tři základní skupiny:

1. *t-contractions* (např. *in it* > *in't; to it* > *to't; knew it* > *knew't*),
2. *the-contractions* (např. *in the* > *i'th/i'th'; of the* > *o'th/o'th'; to the to th'/to'th'*),
3. *s-contractions* (např. *on us/his* > *on's; in us/his* > *in's; make us/his* > *make's*).

Farnham dokládá, že v předpokládané době vzniku *H8* vykazují Shakespearovy, Massingerovy a Fletcherovy hry rozdíly v jejich užívání (TAB. 4.2) a že poměry v *H8* neodpovídají Massingerovu rukopisu, ale podporují Spedding-Hicksonovu atribuci.

	t-contractions	***the-contractions***	***s-contractions***
Shakespeare	velmi časté s množstvím typů	velmi časté	ojedinělé až časté
Massinger	ojedinělé	zcela ojedinělé	nevyskytuje se
Fletcher	časté	časté	ojedinělé

TAB. 4.2: Užívání kontrakcí v dílech Williama Shakespeara, Philipa Massingera a Johna Fletchera v období předpokládaného vzniku hry *Henry VIII* dle Farnham 1916.

Massingerovu účast vrací do hry Henry Dugdale Sykes (1919). Odmítá versologické argumenty s tím, že Massingera a Shakespeara nelze na rovině verše odlišit tak jako dvojici Fletcher – Shakespeare.[33] Sykes postupuje scénu po scéně a dokládá, že v *H8* lze nalézt řadu kolokací, které se objevují i v Massingerově díle, ale v Shakespearově nikoliv, např.: „»We are too open here *to argue this* / Let's think in private more.« Exactly the same thing occurs at the end of one of scenes of *The Bashful Lover* (II. vii.)[…]: »Here's no place / Or time *to argue this*; let us fly hence.« In II. iii. we may note (lines 65-7): […] »More than all my all is nothing.« Compare *The Duke of Milan* I. iii.: […] »All I can pay is nothing.«" (Sykes 1919: 31). Tímto argumentuje pro atribuci, která počítá jen s účastí Fletchera a Massingera (s odlišným rozvržením než Boylova atribuce).

Proti Sykesově atribuci vystoupil Baldwin Maxwell (1926). Dokládá, že řadu kolokací objevujících se později v Massingerově díle obsahuje nejen *H8* ale i starší, prokazatelně Shakespearovy hry. Z toho dovozuje, že výpůjčky ze Shakespeara byly běžnou součástí Massingerova psaní a výskyt paralelních pasáží tak nelze použít jako argument pro jeho autorský podíl na textu *H8*. Dodejme, že Maxwell (1923) zpochybnil i Fletcherovu autorskou účast na *H8* a stavěl se na stranu hypotézy, že jediným autorem je Shakespeare. Proti „metrickým testům" namítá, že četnosti přesahů se v domněle Fletcherových částech *H8* sice vymykají četnostem v Shakespearových hrách, ale jsou

33 „Throughout this paper metrical tests will be altogether disregarded. Such tests are invaluable as a means of distinguishing Fletcher from Shakespeare, but Massinger's metre is so like Shakespeare's that no metrical test has yet been devised to differentiate their work" (Sykes 1919: 22).

významně nižší než ve hrách Fletcherových.[34] Navíc poukazuje na to, že tyto části obsahují řadu pro Fletchera netypických „general thruths“[35] a naopak téměř žádná epizeuxe, která jsou pro Fletchera typická (Maxwell 1923). Tuto hypotézu později podpořil i Peter Alexander (1931), který zpochybnil relevanci „veršových testů“ obecně (naměřené hodnoty přisuzoval posunu v Shakespearově tvůrčím stylu) a proti nim postavil autoritu *Prvního folia*: oba editoři John Heminges a Henry Condell byli Shakespearovými současníky, a je tedy krajně nepravděpodobné, že by pod Shakespearovým jménem publikovali text psaný ve spoluautorství (*nota bene*, že hra *Two Noble Kinsmen*, u níž je Shakespearova a Fletcherova spolupráce téměř jistá, zahrnuta nebyla).

Maxwellova a Alexanderova atribuce podnítila řadu článků, které znovu podpořily Spedding-Hicksonovu atribuci, a to na základě různých textových rysů:

- četnosti veršových přesahů rozlišených podle těsnosti syntaktického vztahu na veršovém švu (Langworthy 1931);
- bohatost slovníku (počet typů slovních tvarů na počet veršů; Hart 1941);
- četnosti variant některých dubletních tvarů (*hath/has, doth/does, them/‘em ye/y‘/you*) a četnosti výskytu slovesa *do* coby expletiva (Partridge 1949);
- četnosti morfologických charakteristik monosylab v ženských zakončeních (Oras 1953);
- hodnoty 10 ukazatelů použitých ve starších studiích naměřených ve Fletcherových a Shakespearových hrách psaných spolu s dalšími autory (Mincoff 1961).[36]

S odlišnými atribucemi přichází Karl Ege (1922), který na základě analýzy četností obrazných pojmenování, slovních paralelismů, aliterací, antitezí a řečnických otázek dochází k závěru, že Shakespearovým spoluautorem nebyl Fletcher, ale jiný dosud neznámý autor, a Cyrus Hoy (1962), který upozorňuje na to, že některé varianty označené Partridgem za typicky fletcherovské nejsou v určitých scénách *H8* distribuovány rovnoměrně, ale hromadí se v relativně krátkých úsecích. Na tomto základě formuluje hypotézu, že původním autorem těchto scén je Shakespeare a Fletcher pouze upravil nebo přidal několik veršů.

Zatímco všechny dosud zmíněné atribuce byly založeny na univariačním přístupu (srovnání izolovaných hodnot), od konce 70. let se i ve výzkumu autorství *H8* začínají prosazovat přístupy multidimenzionální a elementární strojové učení. Jako první se touto cestou vydal Thomas Merriam (1979, 1980), který svou atribuci založil na sadě 20 rysů: četnosti slovních bigramů *by the; in the; it is; to the;* poměry četností *an / a*; *all / (all+any)*; *no / (no+not)*; četnosti výskytu 13 synsémantik typu *and, but, what…* na začátku věty. Merriam ukázal, že rozdíly hodnoty těchto rysů v částech *H8* připisovaných Spedding-Hicksonovou atribucí Shakespearovi a v částech připisovaných Fletcherovi nejsou na obvyklé hladině $\alpha = 0{,}05$ statisticky významné ($\chi^2 = 23{,}85$, $p = 0{,}25$). Merriam pak konstruoval řadu alternativních rozdělení autorství *H8* mezi dva autory a hledal takové, které maximalizuje hodnotu testového kritéria. U výsledného rozdělení ($\chi^2 = 72{,}9$, $p < 10^{-5}$) ukázal, že jedna část vykazuje podobné hodnoty jako Shakespearova *Winter‘s Tale*;

34 Mincoff (1961) upozornil, že Maxwell neprovedl vlastní analýzu, ale opíral se o data uváděná různými autory. Vzhledem k vágnosti termínu „veršový přesah“ tak jeho data mohla být značně nesourodá.

35 „Maxims, proverbs, and concisely worded observations upon human nature“ (Maxwell 1923: 109).

36 Mincoff vycházel z předpokladu, že při spoluautorství mohou být některá specifika autorského stylu potlačena.

druhou část pak rozdělil na tři podskupiny – jednu z nich přisoudil Fletcherovi na základě podobnosti hodnot se vzorky z *The Wild Goose Chase, Bonduca, The Woman's Prize* a *The Elder Brother,* další přisoudil Philipu Massingerovi na základě podobnosti hodnot se vzorky z *The Duke of Milan, The Roman Actor* a *A New Way to Pay Old Debts*; třetí podskupinu přisoudil neznámému autorovi.[37]

Podobně jako Merriam postavil svoji atribuci na četnostech synsémantik i Thomas Horton (1987), který vyčlenil tři sady rysů: (1) četnosti slov *all, dare, in, now, of, the, too, which;* (2) četnosti slov *all, are, in must, now, of, the, two;* (3) četnosti slov *dare, in, of, too, the,* infreq-SH, infreq-FL.[38] Pro každou z těchto sad natrénoval Horton klasifikátor (diskriminační analýza) na prokazatelně Shakespearových a prokazatelně Fletcherových hrách, kterým pak klasifikoval jednotlivé scény *H8* (s tradičním rozdělením druhé scény třetího jednání na dvě části). Hortonova výsledná atribuce – určená jednomyslným hlasováním jednotlivých modelů – ponechala Fletcherovi autorství pouze páté scény pátého jednání, děvět scén přisoudila Shakespearovi a ostatní ponechala nerozhodnuté.

Thomas Merriam přispěl k otázce autorství *H8* řadou dalších článků. Spolu s Robertem Matthewsem (1993) atribuovali několik her včetně *H8* na základě dvou sad rysů: (1) poměry četností *did/(did+do); no/(no+not); to the / to; upon/(on+upon); no/(but+by+for+no+not+so+that+the+to+with)*; (2) relativní četnosti *are, in, no, of, the* (podmnožina rysů užívaných Hortonem (1987)). Pro klasifikaci využili (ve stylometrii jako jedni z prvních vůbec) elementární neuronové sítě, které natrénovali na 10 Shakespearových a 6 Fletcherových hrách. Při atribuci *H8* byly všechny části přisouzeny Shakespearovi. Je ale třeba dodat, že Matthews a Merriam neklasifikovali jednotlivé scény, ale celá jednání (jejichž autorství je podle většiny atribucí smíšené).

Od začátku 21. století se Merriam přiklání k Spedding-Hicksonově atribuci, některé pasáže ovšem reatribuuje zpět Shakespearovi. V článku *Taylor's method applied to Shakespeare and Fletcher* (Merriam 2003a) vychází z konfidenčních intervalů četností slov *all, dare, hath, in, must, sure* a *too* v jednotlivých hrách obsažených v Shakespearově prvním Foliu (vyjma *H8*) a v osmi hrách Johna Fletchera. Ukazuje, že souhrnné četnosti každého ze sedmi slov v částech *H8* připisovaných Spedding-Hicksonovou atribucí Shakespearovi spadají do jejich konfidenčních intervalů v *Prvním Foliu*, zatímco u částí připisovaných Fletcherovi z nich až na jeden případ (*in*) vybočují a jsou blízko hodnotám naměřeným u Fletchera. Vedle toho testuje i vlastní „vylepšenou" („*improved*") atribuci („the beginning and ending of II.ii are Shakespeare's, the Chamberlain's intervention in II.iii is Fletcher's, the beginning of III.i is Shakespeare's, Shakespeare's part of III.ii continues to line 306 rather than 203, the middle of IV.i is Shakespeare's, as is the middle section of IV.ii." (Merriam 2003a: 422); podrobnosti viz OBR. 4.1) – v tomto případě vybočují v částech připisovaných Fletcherovi z konfidenčního intervalu Prvního Folia četnosti všech sedmi slov. Ve prospěch nové atribuce pak argumentuje ještě analýzou hlavních komponent, kde se Fletcherovy části *H8* ocitají blíž hrám Johna Fletchera než části připisované Fletcherovi Spedding-Hicksonovou atribucí.[39]

37 V případě *Winter's Tale* uvádí Merriam jako měřítko podobnosti hodnotu testového kritéria χ^2, u ostatních mluví jen neurčitě o „statistical resemblance" / „characterized by similarities".

38 Infreq-SH značí souhrnnou četnost slov s nízkou frekvencí typických pro Shakespeara, infreq-FL souhrnnou četnost slov s nízkou frekvencí typických pro Fletchera.

39 Viz též převážně interpretačně zaměřená publikace *The Identity of Shakespeare in Henry VIII* (Merriam 2005).

Následně (2003b, 2014) se hlavním argumentem pro Merriamovu atribuci stává kontroverzní technika CUSUM.[40] Merriam vymezuje nový soubor 22 relevantních textových rysů (výskyt *all, are, conscience, did, 'em, find, from, hath, in, is, it, little, must, now, sure, they, 'tis, too, where, there,* výskyt slov zakončených na *-ly* a výskyt ženský zakončených veršů), které zpracovává metodou kumulativní sumace (CUSUM), která jednoduchým způsobem vyhodnocuje nepravidelnosti v jejich distribuci napříč textem: u každého *i*-tého verše z celkového počtu *n* je spočten celkový počet jevů z výše definované množiny (m_i) a od této hodnoty je odečten průměrný počet na jeden verš:

$$r_i = m_i - \sum_{j=1}^{n} \frac{m_j}{n} \qquad (4.1)$$

Výsledná hodnota q_i je pak pro každý *i*-tý verš určena jako součet hodnoty r_i a všech *r* předcházejících veršů:

$$q_i = \sum_{j=1}^{i} r_j \qquad (4.2),$$

Změny v průběhu křivky tvořené vynesením jednotlivých hodnot $r_i \in \langle 0; n \rangle$ Merriam vysvětluje změnami autorství v rámci jednotlivých scén (od Merriam 2005 rozděluje oproti 2003a, 2003b druhou scénu třetího jednání odlišně; viz OBR. 4.1).[41]

V zatím posledním článku Merriam (2018) podpořil svoji atribuci analýzou hlavních komponent frekvencí 64 nejčetnějších slov.

Mark Eisen, Alejandro Riberio, Santiago Segarra a Gabriel Egan (2017) atribuovali jednotlivé scény *H8* na základě četnosti výskytů kolokací vybraných synsémantik pomocí *Word Adjacency Networks* (Segarra et al. 2013). Oproti Spedding-Hicksonově atribuci reatribuují některé scény zpět Shakespearovi.

Vedle toho se v průběhu devadesátých, nultých a desátých let objevují i články založené na tradičním univariačním přístupu. Kromě již zmiňovaných versologických atribucí Mariny Tarlinské (1987, 2014) patří do této skupiny např. Jonathan Hope (1994/2009) nebo MacDonnald P. Jackson (1997). Hope rozdělil text *H8* na tři části podle vlivného článku Cyruse Hoye (1962): (A) Shakespearovy texty, (B) Fletcherovy texty, (C) Shakespearovy texty s Fletcherovými úpravami nebo dodatky. U těchto tří částí srovnal (1) četnost *do* coby expletiva, (2) poměry četností podřadicích konektorů vedlejších vztažných vět (*who,*

40 Metodu CUSUM zpopularizoval na začátku 90. let zejm. Andrew Q. Morton. Přestože mnozí odborníci její úspěšnost od počátku zpochybňovali, byla coby forenzní technika několikrát užita u britských soudů. K pádu z výsluní došlo v roce 1993, kdy Morton v živém přenosu britské televize nebyl schopen rozlišit mezi texty napsanými usvědčeným zločincem a texty napsanými předsedou nejvyššího soudu (podrobněji Grieve 2005: 47–49; Juola 2006: 243–244). Merriam ovšem podotýká, že technika CUSUM byla zavedena a s úspěchem používána dlouho před Mortonem (Woodward–Goldsmith 1964) a že za jeho selháním stojí především nedostatečný soubor analyzovaných rysů: „Serious criticisms of Morton and Farringdon are based on their selection of such variables as word and sentence length from narrow text samples, and the reliability of the assumed consistent relationship between two or three such variables within single authors as determined by the superimposition of cusum graphs. These criticisms of the 'Qsum' method are not directed at the fundamental technique introduced by Woodward and Goldsmith." (Merriam 2000: 163).

41 Proti této atribuci vystoupil MacDonnald P. Jackson (2013) s poukazem na to, že pasáže, které reatribuuje Shakespearovi obsahují řadu tradičních Fletcherovských markerů. Reakce viz Merriam 2014.

which, that, ∅), (3) poměr četností *you* a *thou*. Na tomto základě podpořil proti Hoyovi Spedding-Hicksonovu atribuci. Jackson dospěl ke stejnému výsledku srovnáním délky frází měřené počtem slov v jednotlivých scénách.

Dodejme, že i v dnešní době se objevují názory, že jediným autorem *H8* je Shakespeare (srov. Vickers 2004).

***

Rozsáhlý výčet měl za cíl ukázat, že vedle kanonické Spedding-Hicksonovy atribuce existuje celá řada alternativních hypotéz a že otázku autorství *H8* nelze stále považovat za uzavřenou.

4.1.2 Data

Pro potřeby atribuce *H8* byl využit korpus starší anglické literatury *Earlyprint* budovaný na půdě Northwestern University a Washington University obsahující mimo jiné i původní vydání renesančních dramat (<https://drama.earlyprint.org>). Původní korpus je tokenizovaný (přičemž ke každému slovnímu tvaru je připojeno i jeho znění v současné angličtině – „standardizovaný slovní tvar“), lemmatizovaný a morfologicky označkovaný (s rozsáhlými manuálními úpravami), anotace přízvuků a rozpoznávání veršových rozměrů bylo provedeno dodatečně pomocí balíčku *Prosodic* (Antilla–Heuser 2016).

Jako trénovací data byly využity jednotlivé scény her autorů zmiňovaných ve výše shrnutých článcích (kap. 4.1.1) jako možní kandidáti – Williama Shakespeara, Johna Fletchera a Philipa Massingera – pocházející zhruba z období předpokládaného vzniku *H8*, konkrétně:

(1) Shakespeare: *The Tragedy of Coriolanus* (5 scén)

(2) Shakespeare: *The Tragedy of Cymbeline* (27 scén)

(3) Shakespeare: *The Winter‘s Tale* (12 scén)

(4) Shakespeare: *The Tempest* (9 scén)

(5) Fletcher: *Valentinian* (21 scén)

(6) Fletcher: *Monsieur Thomas* (28 scén)

(7) Fletcher: *The Woman‘s Prize* (23 scén)

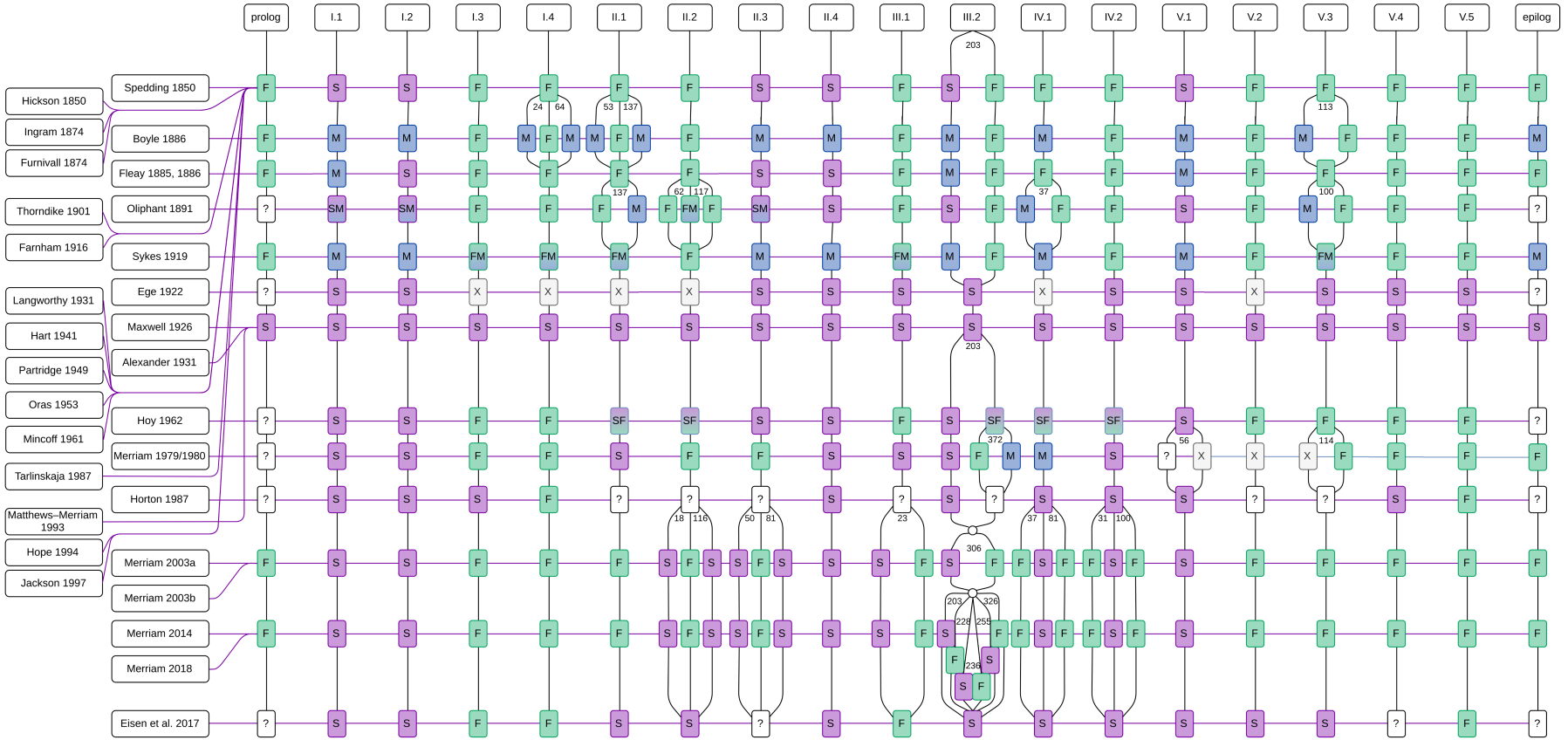
(8) Fletcher: *Bonduca* (18 scén)

(9) Massinger: *The Duke of Milan* (10 scén)

(10) Massinger: *The Unnatural Combat* (11 scén)

(11) Massinger: *The Renegado* (25 scén)

Celkem tak bylo k dispozici 53 trénovacích vzorků pro Shakespeara, 90 trénovacích vzorků pro Fletchera a 46 trénovacích vzorků pro Massingera.



OBR. 4.1: Přehled atribucí hry *Henry VIII.* S = William Shakespeare, F = John Fletcher, M = Philip Massinger, X = jiný, neznámý autor, ? = neurčeno. U scén se smíšeným autorstvím, kde atribuce neudává konkrétní rozdělení: FM = Fletcher/Massinger, SM = Shakespeare/Masinger, SF = Shakespeare/Fletcher. U scén se smíšeným autorstvím, kde atribuce udává konkrétní rozdělení, označují údaje mezi vertikálními spojnicemi číslo verše, od něhož začíná část připisovaná následujícímu autorovi (číslování veršů dle Pooler, C. K. (ed.) (1915). The Famous History of the Life of King Henry VIII. London: The Arden Shakespeare).

U všech textů bylo zpracováno jejich první vydání. U zmiňovaných Shakespearových her, stejně jako u samotného *H8*, bylo zpracováno vydání z *Prvního folia* (J. Heminges – H. Condell (eds.) (1623). *Mr. VViliam Shakespeares Comedies, Histories, & Tragedies*). U Fletcherových her bylo použito: (5), (7), (8) *Beaumontovo a Fletcherovo první folio* (H. Moseley – H. Robinson (eds.) (1647). *Comedies and Tragedies Written by Francis Beavmont and John Fletcher Gentlemen*), u (6) samostatné vydání (T. Harper (ed.) (1639). *Monsievr Thomas*). U Massingerových her: (9) E. Blackmore (ed.) (1623). *The Dvke of Millaine*, (10) J. Waterson (ed.) (1639). *The Vnnatvrall Combat*, (11) J. Waterson (ed.) (1630). *The Renegado, A Tragae Comedie*.

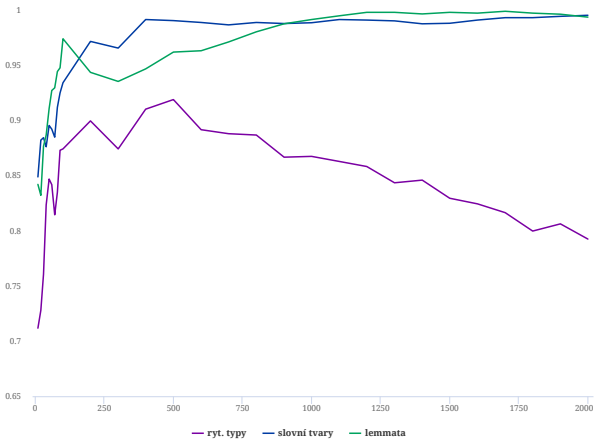
4.1.3 Atribuce scén pomocí SVM

V prvním kroku byla na trénovacích vzorcích testována úspěšnost rozpoznávání na základě jednotlivých rysů. Abychom předešli riziku overfittingu, byly vzorky při křížové validaci vždy rozděleny tak, aby testovací množina obsahovala všechny scény jedné hry, tzn. nejprve bylo klasifikováno 5 scén Shakespearova *Coriolana* pomocí modelu natrénovaného na zbylých třech Shakespearových hrách (2–4), čtyřech Fletcherových hrách (5–8) a třech Massingerových hrách (9–11), poté bylo klasifikováno 27 scén ze hry *Cymbeline* pomocí modelu natrénovaného na hrách 1 a 2–11 atd. Vzhledem k tomu, že počty vzorků (scén) se u každého autora liší, bylo jejich množství při každém kroku křížové validace zarovnáno náhodným výběrem. Celý proces byl pak – stejně jako u předchozích experimentů – 30krát opakován.

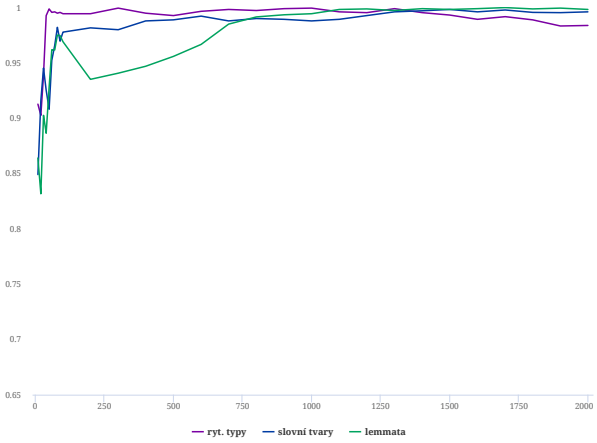
Při pilotním testování se ukázalo, že při stejném nastavení vykazují četnosti slovních tvarů a znakových *n*-gramů znatelně nižší úspěšnost než četnosti lemmat a standardizovaných slovních tvarů. (Příčinou je pravděpodobně variabilita pravopisu napříč jednotlivými editory textů.) Z rysů modelujících veršový rytmus se jako nejspolehlivější ukázaly rytmické typy (srov. kap. 2.1.2).[42]

OBR. 4.2 ukazuje výsledky křížových validací pro n nejčetnějších rytmických typů, standardizovaných slovních tvarů (dále jen „slovní tvary“) a lemmat ($n \in \{10, 20, 30, \ldots, 100, 200, 300, \ldots, 2000\}$). Každá ze tří sad rysů se ukazuje jako spolehlivý ukazatel autorství: slovní tvary dosahují maxima úspěšnosti (~0,992) při hodnotě n = 400, přičemž úspěšnost zůstává nadále poměrně stabilní; úspěšnost lemmat je zpočátku nižší a maxima dosahuje až při hodnotě n = 1200 (~0,997); rytmické typy dosahují maxima při hodnotě n = 500 (~0,919), poté jejich úspěšnost významně klesá. Jiný obrázek ale dostaneme, když se podíváme na výsledky rozpoznávání pouze dvojic autorů. Při vypuštění Massingerových her, tj. rozpoznávání Shakespeara a Fletchera (OBR. 4.3) dosahují rytmické typy maxima při n = 50 (~0,999) a jejich úspěšnost zůstává i nadále poměrně stabilní; lemmata a slovní tvary dosahují těchto hodnot až při n = 1100, resp. n = 1300. Při vypuštění Fletcherových her, tj. rozpoznávání Shakespeara a Massingera (OBR. 4.4) dosahují rytmické typy maxima (~0,954) až při hodnotě n = 400 a následně jejich úspěšnost klesá. Významně tak zaostávají za slovními tvary i lemmaty, u nichž je už od hodnoty n = 400 úspěšnost stoprocentní. Do určité míry se tak potvrzuje teze Baldwina Maxwella (viz pozn. 33), že zatímco Fletcherův rytmický styl se od Shakespearova značně odlišuje, Massingerův je mu výrazně podobný.

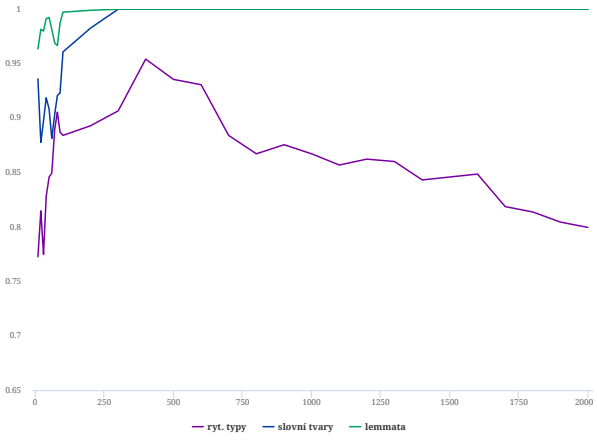
42 Analyzovány byly rytmické typy všech veršových rozměrů. Jednotlivé části stichomytie (jeden verš rozdělený mezi více postav a vysázený na více řádků) byly brány jako samostatné verše – takový přístup se při testování ukázal jako produktivnější než jejich spojování do jednoho verše.



OBR. 4.2: Výsledky křížových validací Shakespearových, Fletcherových a Massingerových vzorků pro n nejčetnějších rytmických typů, standardizovaných slovních tvarů a lemmat ($n \in \{10, 20, 30, \ldots, 100, 200, 300, \ldots, 2000\}$).



OBR. 4.3: Výsledky křížových validací Shakespearových a Fletcherových vzorků pro n nejčetnějších rytmických typů, standardizovaných slovních tvarů a lemmat ($n \in \{10, 20, 30, \ldots, 100, 200, 300, \ldots, 2000\}$).



OBR. 4.4: Výsledky křížových validací Shakespearových a Massingerových vzorků pro n nejčetnějších rytmických typů, standardizovaných slovních tvarů a lemmat ($n \in \{10, 20, 30, \ldots, 100, 200, 300, \ldots, 2000\}$).

S ohledem na výsledky křížových validací byla atribuce *H8* provedena pomocí kombinovaného modelu založeného na 500 nejčetnějších slovních tvarech a 500 nejčetnějších rytmických typech (slovní tvary byly upřednostněny před lemmaty proto, že vykazují srovnatelnou úspěšnost při mnohem nižších hodnotách *n*, hladina *n* = 500 představuje hranici, po níž už v žádné ze tří sérií validací nedocházelo k významnému nárůstu úspěšnosti a zároveň ani k její degradaci). TAB. 4.3 ukazuje, že kombinovaný model je při klasifikaci her obsažených v trénovacích datech vždy minimálně tak úspěšný jako silnější z dílčích klasifikátorů (slovní tvary), ve třech případech ještě lehce úspěšnější.

		ryt. typy	slovní tvary	kombinace
Shakespeare	*Coriolanus*	0,98	1	1
	Cymbeline	0,98	1	1
	Winter's Tale	0,99	1	1
	Tempest	0,97	1	1
Fletcher	*Valentinian*	0,84	0,95	0,96
	Monsieur Thomas	1	0,98	1
	Woman's Prize	0,98	1	1
	Bonduca	0,98	0,98	1
Massinger	*Duke of Milan*	0,81	0,99	0,99
	Unnatural Combat	0,83	1	1
	Renegado	0,88	1	1

TAB. 4.3: Úspěšnost rozpoznávání autorství scén jednotlivých her pomocí SVM modelů založených na (1) četnostech 500 nejfrekventovanějších rytmických typů, (2) četnostech 500 nejfrekventovanějších slovních tvarů, (3) 1000rozměrných vektorech kombinujících rysy (1) a (2).

TAB. 4.4 zobrazuje výsledky klasifikací jednotlivých scén *H8* pomocí kombinovaných klasifikátorů[43] – číslo udává v kolika případech z 30 klasifikací byla scéna atribuována danému autorovi. Z výsledků lze dovozovat:

(1) Massingerova účast na *H8* je nepravděpodobná. Ze 17 scén byly některým z 30 modelů jako Massingerovy klasifikovány pouze dvě (2.1, 4.2), v obou případech šlo navíc o minoritu modelů.

(2) To, že text vznikl ve spoluautorství Shakespeara a Fletchera, je vysoce pravděpodobné: u sedmi scén predikuje všech 30 modelů jednomyslně Shakespearovo autorství, u pěti scén predikuje všech 30 modelů jednomyslně Fletcherovo autorství.

(3) Výsledky do značné míry korespondují se Spedding-Hicksonovou atribucí. Kromě dvou případů predikuje vždy majorita modelů jí přisuzuovaného autora; výjimkou je třetí scéna druhého jednání, u níž Spedding-Hicksonova atribuce předpokládá smíšené autorství, a první scéna čtvrtého jednání, kterou přisuzuje Fletcherovi.

Tato atribuce samozřejmě stojí na implicitním předpokladu, že rozdělení autorství mezi Williama Shakespeara a Johna Fletchera koresponduje s rozdělením scén, který, jak jsme viděli výše (kap. 4.1.1), nebývá vždy bez výhrad přijímán (Hoy 1962; Merriam 2003a, 2003b, 2014, 2018). V dalších dvou kapitolách se proto zaměříme na dvě techniky, které dělení na scény neberou v potaz: (1) výše popsaná metoda CUSUM aplikovaná Thomasem Merriamem (kap. 4.1.4) a (2) technika „klouzavé atribuce" navrhovaná Maciejem Ederem (kap. 4.1.5).

43 Prolog, epilog a druhá scéna pátého jednání (obsahující převážně prózu) nebyly klasifikovány kvůli nízkému počtu veršů.

	Výsledky klasifikací			
	Shakespeare	**Fletcher**	**Massinger**	**Spedding-Hicksonova atribuce**
1.1	**30**	0	0	Shakespeare
1.2	**30**	0	0	Shakespeare
1.3	0	**30**	0	Fletcher
1.4	0	**30**	0	Fletcher
2.1	0	**20**	10	Fletcher
2.2	0	**30**	0	Fletcher
2.3	**30**	0	0	Shakespeare
2.4	**30**	0	0	Shakespeare
3.1	0	**30**	0	Fletcher
3.2	**30**	0	0	**Shakespeare/Fletcher**
4.1	**30**	0	0	**Fletcher**
4.2	0	**23**	7	Fletcher
5.1	**30**	0	0	Shakespeare
5.3	9	**21**	0	Fletcher
5.4	7	**23**	0	Fletcher
5.5	0	**30**	0	Fletcher

TAB. 4.4: Výsledky klasifikací jednotlivých scén H8 – číslo udává, v kolika případech z 30 klasifikací byl scéně predikován daný autor. Nejvyšší hodnota v každém řádku je zvýrazněna tučně. Sloupec vpravo udává, komu přisuzuje autorství Spedding-Hicksonova atribuce; rozdíly oproti našim výsledkům jsou zvýrazněny tučně.

4.1.4 Thomas Merriam: Atribuce na základě CUSUM (Cumulative Sum)

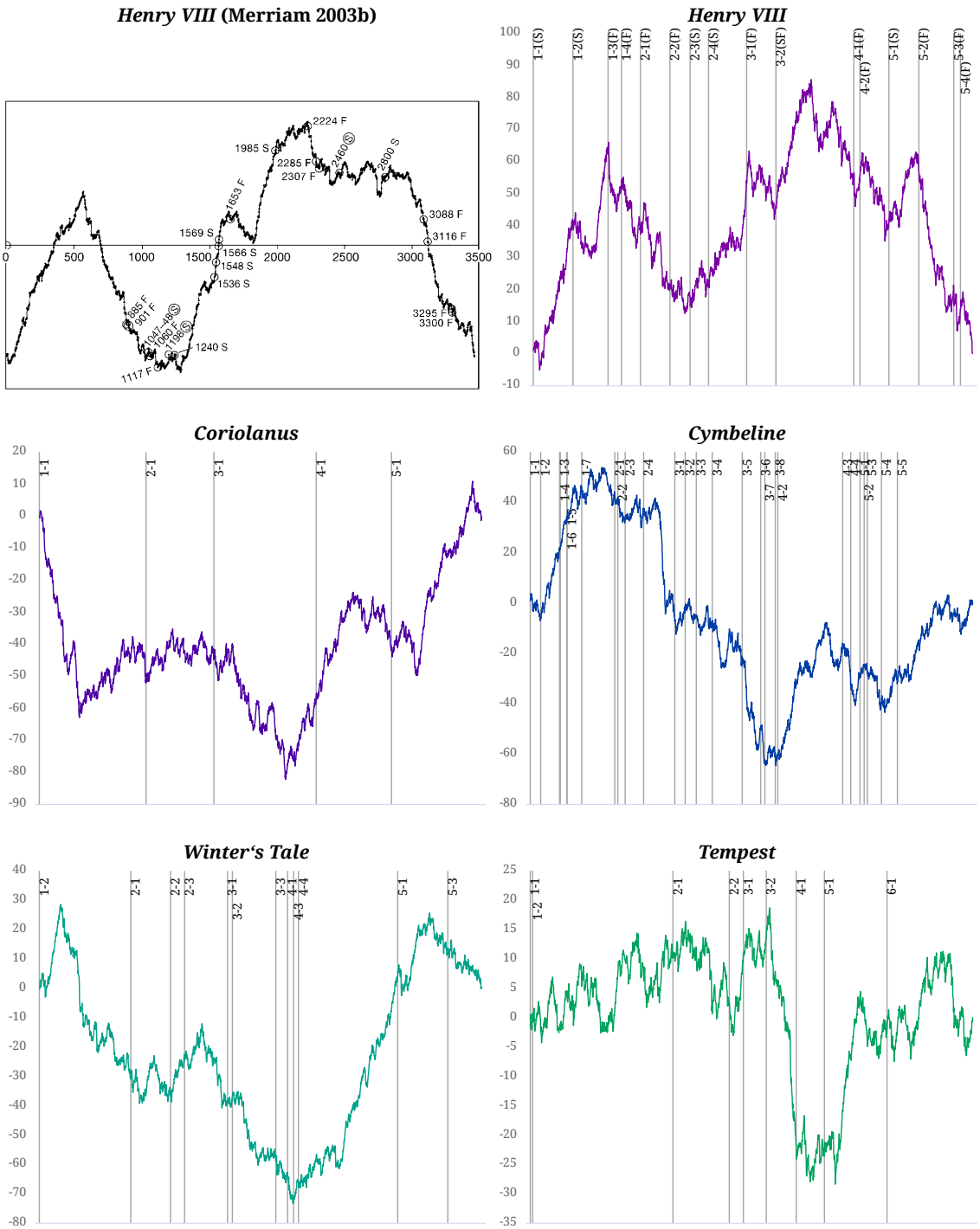
Vzhledem k tomu, že Thomas Merriam výše popsanou analýzu metodou CUSUM označuje za první a dosud jedinou analýzu autorství hry *H8*, která nebere v potaz dělení na scény,[44] pokusili jsme se jeho výsledky reprodukovat a zároveň otestovat její úspěšnost na čtyřech Shakespearových hrách obsažených v našem trénovacím korpusu.

OBR. 4.5 obsahuje přetisk Merriamova původního grafu (Merriam 2003b) zobrazujícího pro *H8* křivku CUSUM konstruovanou na základě 22 relevantních textových rysů (výskyt *all, are, conscience, did, 'em, find, from, hath, in, is, it, little, must, now, sure, they, 'tis, too, where, there,* výskyt slov zakončených na *-ly* a výskyt žensky zakončených veršů) a stejným způsobem námi konstruovanou křivku pro hry *H8, Coriolanus, Cymbeline, Winter's Tale* a *Tempest*.

Je zřejmé, že Merriamova křivka se s naší křivkou konstruovanou pro *H8* v hlavních trendech shoduje (dílčí odlišnosti lze patrně přičítat zejm. odlišnému traktování stichomythií; srov. pozn. 42 výše). U námi konstruované křivky můžeme pozorovat, že poměrně jednoznačně identifikuje většinu změn autorství předpokládaných Spedding-Hicksonovou atribucí:

- sestupná tendence začínající na hranicích scén 1.2/1.3 indikuje začátek Fletcherových pasáží (po prvních dvou Shakespearových scénách);
- vzestupná tendence začínající na hranicích scén 2.2/2.3 indikuje začátek Shakespearových pasáží;
- sestupná tendence začínající na hranicích scén 2.4/3.1 indikuje začátek Shakespearových pasáží;

44 „I know of no other such line-by-line analysis of the authorship of *Henry VIII* to date" (Merriam 2003b: 425).



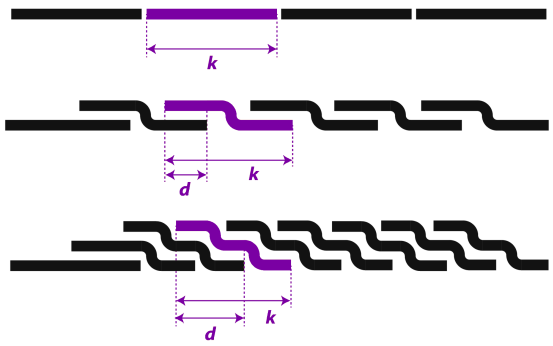
OBR. 4.5: Kumulativní sumace (Cusum) her *H8* (faksimile z Merriam 2003b a vlastní výpočet), *Coriolanus, Cymbeline, Winter's Tale, Tempest* založená na 21 textových rysech. Svislé čáry vyznačují hranice mezi scénami. U *H8* udává symbol v závorce, komu přisuzuje autorství Spedding-Hicksonova atribuce (S: Shakespeare, F: Fletcher, SF: smíšené autorství).

- vzestup a následný pokles křivky v rámci scény 3.2 indikuje smíšené autorství této scény v pořadí Shakespeare – Fletcher;
- následná sestupná tendence křivky narušená vzestupem ve scéně 5.1 indikuje, že autorem zbytku hry je – s výjimkou zmíněné scény – Fletcher.

Problémem ovšem je, že srovnatelně výrazné změny trendu můžeme nalézt i u křivek konstruovaných pro ostatní hry, u nichž není důvod předpokládat autorskou účast někoho jiného než Williama Shakespeara. Můžeme předpokládat, že za těmito změnami stojí zejm. tematické a motivické přechody a že tedy metoda CUSUM s sebou nese vysoké riziko falešně pozitivních výsledků.

4.1.5 Klouzavá atribuce pomocí SVM

Klouzavou atribuci (*rolling attribution*) zavádí Maciej Eder (2016) jako atribuční techniku pro případy smíšeného autorství. Na rozdíl od běžných úloh není atribuován celý text, resp. jeho logické části (kapitoly, scény...), ale navzájem se překrývající úseky. U textu, kde předpokládáme spolupráci dvou (příp. více) autorů, sestávajícího z n veršů l_1, l_2, l_3, … l_n jsou arbitrárně zvolené hodnoty k; $k \in \mathbb{N}$, $k < n$ a d; $d \in \mathbb{N}$, $d < n - k$, $d \leq k$ a následně pro všechna i; $i \in \{0, d, 2d, 3d, ...\}$, $i < n - k$ provedena sada atribucí všech úseků s_i sestávajících z veršů l_{i+1}, l_{i+2}, l_{i+3}, …, l_{i+k}. (viz OBR. 4.6).[45] Pro větší citlivost ke změnám autorství doporučuje Eder pracovat ne s prostými predikcemi (příslušnost úseku k jedné ze dvou, příp. více tříd), ale – pokud to zvolený klasifikátor umožňuje – s rozdělením pravděpodobnosti mezi dané třídy.



OBR. 4.6: Schematické znázornění principu klouzavé atribuce. Nahoře: $d = k$, uprostřed: $d > k / 2$, dole: $d < k / 2$. Zdroj: Eder 2016 (faksimile).

Vzhledem k tomu, že autorskou účast Philipa Massingera jsme výše (kap. 4.1.3) vyhodnotili jako velmi nepravděpodobnou, omezili jsme u klouzavé atribuce kandidátskou množinu na Williama Shakespeara a Johna Fletchera. Text *H8* byl segmentován na 569 překrývajících se úseků při k = 100 a d = 5. Celý proces byl opět 30krát zopakován s novým náhodným zarovnáním počtu trénovacích vzorků. Každé pětiverší (vyjma 19 počátečních a 19 závěrečných) tak bylo klasifikováno celkem 600krát (30krát v rámci 20 různých úseků). Pro odhad pravděpodobností z výsledků klasifikace pomocí SVM bylo využito Plattovo škálování (*Platt's scaling*; Platt 1999).

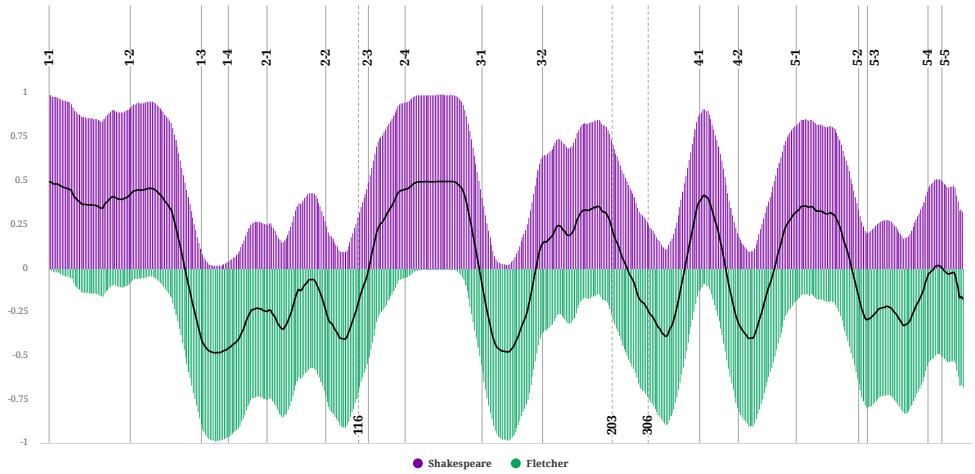
45 V původním Ederově návrhu je mírou délky úseků počet slov.

OBR. 4.7, 4.8 a 4.9 zobrazují výsledky klouzavé atribuce na základě (1) 500 nejčetnějších rytmických typů, (2) 500 nejčetnějších slovních tvarů a (3) 1000rozměrných vektorů kombinujících rysy (1) a (2). Každý sloupec odpovídá jednomu pětiverší a zobrazuje průměr pravděpodobností Shakespearova a Fletcherova autorství, které mu byly v rámci různých úseků přiděleny při 30 iteracích. Pro lepší přehlednost jsou hodnoty u Fletchera zobrazeny jako záporné (vzdálenost mezi vrcholy Shakespearova a Fletcherova sloupce je tak vždy rovna jedné). Černá křivka udává průměr obou zobrazených hodnot. Svislé čáry ukazují jednak hranice mezi scénami (plná čára), jednak některá výše (kap. 4.1.1) zmiňovaná místa v textu (přerušovaná čára, popisek udává číslo verše v rámci dané scény).[46]

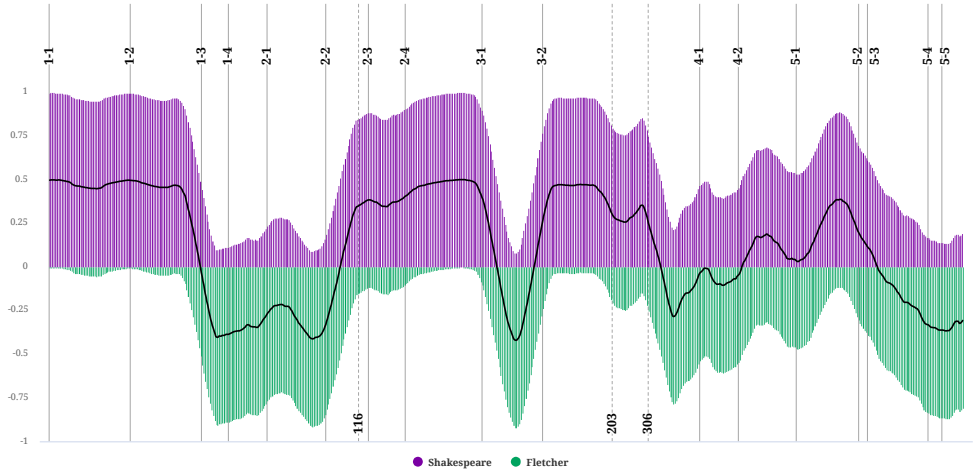
I výsledky atribuce nezohledňující dělení na scény mluví do značné míry ve prospěch Spedding-Hicksonovy atribuce:

- U scén **1.1** a **1.2** ukazují rytmické typy, slovní tvary i kombinovaný model na velmi vysokou pravděpodobnost Shakespearova autorství; pravděpodobné místo změny autorství se u všech tří modelů kryje s koncem scény 1.2.
- U scén **1.3, 1.4, 2.1** a **2.2** ukazují všechny tři typy modelů na vysokou pravděpodobnost Fletcherova autorství; u modelu založeného na četnostech rytmických typů se pravděpodobné místo změny autorství kryje s koncem scény 2.2, u modelu založeného na četnostech slovních tvarů došlo ke změně autorství pravděpodobně už během této scény (změnu autorství – byť o něco později (verš 116) – zde předpokládají i atribuce Thomase Merriama; srov. OBR. 3.3 výše).
- Scény **2.3** a **2.4** jsou dle všech tří modelů s velmi vysokou pravděpodobností Shakespearovy; pravděpodobné místo změny autorství se u všech tří modelů kryje s koncem scény 2.4.
- Scéna **3.1** je dle všech tří modelů s vysokou pravděpodobností Fletcherova; pravděpodobné místo změny autorství se u všech tří modelů kryje s koncem této scény.
- U scény **3.2** se tradičně předpokládá smíšené autorství v pořadí Shakespeare – Fletcher. Výsledky všech tří modelů tomuto předpokladu odpovídají. Zatímco ale Spedding-Hicksonova atribuce předpokládá, že Fletcherova část začíná veršem 203 (odchod krále ze scény), všechny tři modely indikují změnu autorství později. Kombinovaný model lokalizuje změnu přesně u verše 306, kde ji předpokládají i starší atribuce Thomase Merriama (2003a, 2003b). Určitý pokles pravděpodobnosti Shakespearova autorství v okolí verše 306 u modelu založeného na četnostech slovních tvarů i kombinovaného modelu by mohl svědčit ve prospěch pozdějších Merriamových atribucí (2014, 2018 – smíšené autorství i po verši 203).
- U scén **4.1** a **4.2** ukazuje model založený na četnostech rytmických typů na pravděpodobné Shakespearovo autorství u první z nich (v rozporu se Spedding-Hicksonovou atribucí) a Fletcherovo autorství u druhé; předpokládané změny autorství se ale zcela nekryjí s hranicemi scén. Pravděpodobnosti dovozené z modelu založeného na četnostech slovních tvarů jsou u těchto scén pro oba autory blízko hodnotě 0,5, což by mohlo svědčit ve prospěch atribucí Thomase Merriama (smíšené autorství).

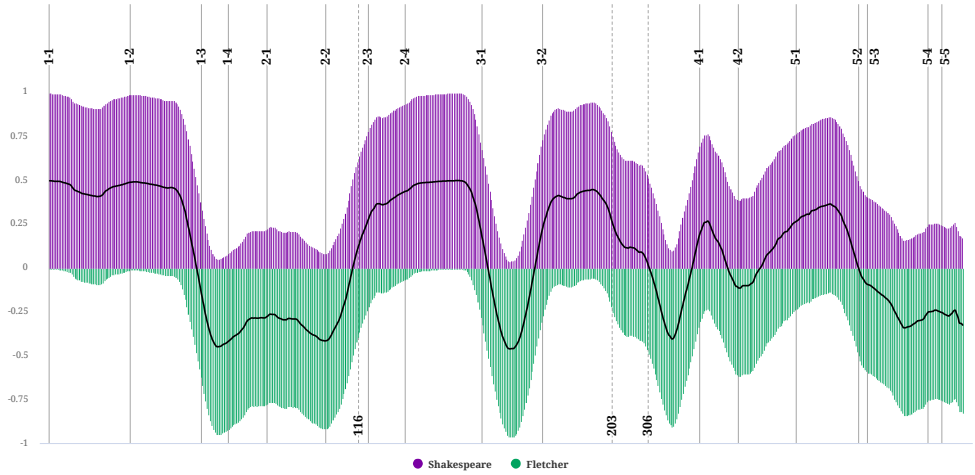
46 Číslování veršů dle Pooler, C. K. (ed.) (1915). *The Famous History of the Life of King Henry VIII*. London: The Arden Shakespeare.



OBR. 4.7: Klouzavá atribuce *H8* na základě 500 nejčetnějších rytmických typů.



OBR. 4.8: Klouzavá atribuce *H8* na základě 500 nejčetnějších slovních tvarů.



OBR. 4.9: Klouzavá atribuce *H8* na základě 500 nejčetnějších rytmických typů a 500 nejčetnějších slovních tvarů.

- Scéna **5.1** je dle všech tří modelů s vysokou pravděpodobností Shakespearova; pravděpodobné místo změny autorství se u modelu založeného na rytmických typech i u kombinovaného modelu kryje s koncem této scény, dle modelu založeného na slovních tvarech dochází ke změně autorství o něco později.
- Scény **5.2**, **5.3**, **5.4** a **5.5** jsou dle modelu založeného na četnostech slovních tvarů a kombinovaného modelu s vysokou pravděpodobností Fletcherovy. Model založený na četnostech rytmických typů naznačuje možnou Shakespearovu účast na scéně **5.4**.

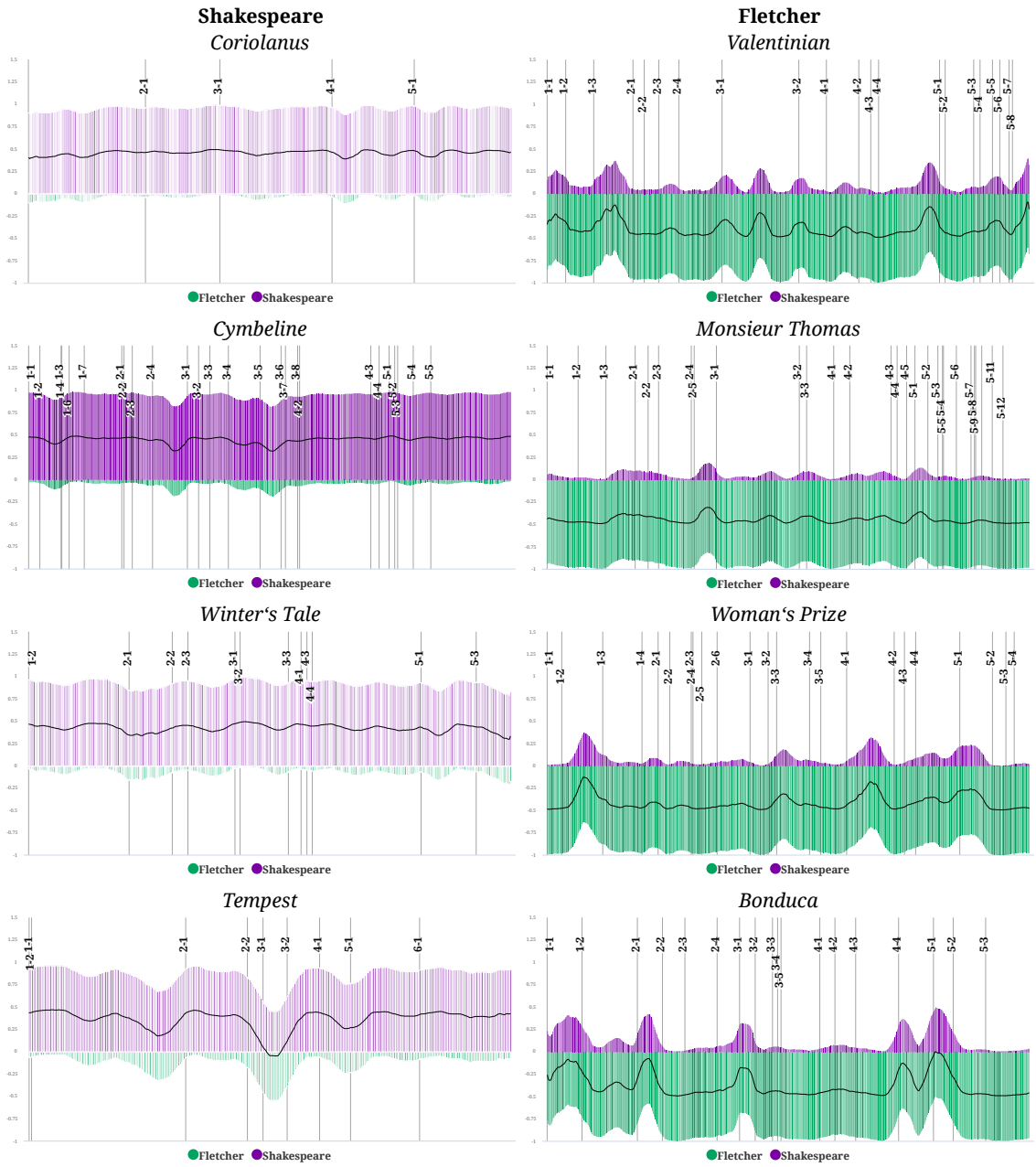
Pro ilustraci uvádíme výsledky aplikace techniky klouzavé atribuce (se stejným nastavením jako u *H8* výše) pomocí kombinovaného modelu na čtyři Shakespearovy a čtyři Fletcherovy hry obsažené v trénovacím korpusu (OBR. 4.10). Na rozdíl od Merriamovy aplikace metody CUSUM jsou výsledky aplikace klouzavé atribuce v souladu s očekáváním: (1) pravděpodobnost Fletcherova autorství je u velké většiny úseků Shakespearových her minimální, vyšších hodnot než Shakespeare dosahuje Fletcher pouze u deseti pětiverší ve třetí scéně druhého jednání hry *Tempest*, (2) pravděpodobnost Shakespearova autorství je u velké většiny úseků Fletcherových her minimální, Fletcherovým hodnotám se Shakespeare blíží pouze u první scény pátého jednání hry *Bonduca*. Při 10 chybně atribuovaných pětiverších z celkového počtu 4412 tak můžeme úspěšnost metody klouzavé atribuce odhadovat na 0,9977 a výše provedenou analýzu *H8* označit za vysoce spolehlivou.

4.1.6 Shrnutí

Kombinované versologicko-lexikální modely natrénované na anglických dramatech z počátku 17. století vykazují vysokou úspěšnost rozpoznávání autorství. S vysokou mírou spolehlivosti tak lze na základě provedených analýz tvrdit, že na veršovaném dramatu *The Famous History of the Life of King Henry the Eight* se autorsky podílel jak William Shakespeare, tak John Fletcher, zatímco autorská účast Philipa Massingera je velmi nepravděpodobná.

Metoda klouzavé atribuce (*rolling attribution*) ukázala, že rozdělení autorství mezi Shakespeara a Fletchera většinou koresponduje s dělením textu na jednotlivé scény, a to do značné míry v souladu se Spedding-Hicksonovou atribucí (včetně předpokládaného smíšeného autorství druhé scény třetího jednání). Hlavní rozdíl oproti Spedding-Hicksonově atribuci představují u našich modelů ambivalentní výsledky u obou scén čtvrtého jednání. Dodejme ale, že už sám James Spedding vyslovil ohledně atribuce těchto scén Fletcherovi určité pochyby.[47] Ostatní rozdíly jsou spíše marginální; buď svědčí ve prospěch drobných úprav, které vůči Spedding-Hicksonově atribuci navrhuje Thomas Merriam, nebo pocházejí ze situací, kdy si výsledky jednotlivých modelů odporují. Jejich interpretaci ponechme povolanějším.

47 „Of the 4th Act I did not so well know what to think. For the most part it seemed to bear evidence of a more vigorous hand than Fletcher's, with less mannerism, especially in the description of the coronation, and the character of Wolsey; and yet it had not to my mind the freshness and originality of Shakspere" (Spedding 1850: 119).



OBR. 4.10: Klouzavá atribuce Shakespearových her *Coriolanus*, *Cymbeline, Winter's Tale* a *Tempest* a Fletcherových her *Valentinian, Monsieur Thomas, Woman's Prize* a *Bonduca* na základě 500 nejčetnějších rytmických typů a 500 nejčetnějších slovních tvarů.

4.2 Autorství básní připisovaných Josefu Barákovi

V roce 1958 publikoval Oldřich Králík pod názvem *Z doby Májů* kompletní edici básní podepsaných Josefem Barákem (původně otištěných mezi lety 1858–1862 ve sbornících a časopisech). K nim dále připojil i povídku *Kříž pod Petřínem* publikovanou pod Barákovým jménem v *Almanachu Máj*. Na obálce ani titulním listě ovšem Barákovo jméno neuvedl. V té době už se totiž Králík několik let přikláněl k hypotéze, že skutečným autorem těchto textů není Josef Barák, ale jeho přítel Jan Neruda.

Králík už v roce 1956 otiskl v časopise *Host do domu* povídku *Kříž pod Petřínem* s uvedením Nerudy coby autora, což zdůvodňoval v komentáři (Králík 1956) jednak svědectvím Nerudovy přítelkyně Anny Holinové, jednak tematickou příbuzností a podobností lokace (Malá Strana) s Nerudovou povídkovou tvorbou. O rok později rozšířil Králík hypotézu Nerudova autorství i na básnické texty. K tomu dodává: „Nejde zatím ovšem o nic víc než o hypotézu – byť byla sebepravděpodobnější“ (1957: 6). V komentáři k výše zmíněné edici už ale tuto variantu pokládá za *de facto* jistou: „To, co víme o Barákovi z jeho vlastních vzpomínek a odjinud, málo podporuje představu, že by byl schopen napsat tak výjimečné básně a tak odvážnou povídku. Naopak mnoho okolností nasvědčuje domněnce, že za Barákem se skrývá jeho důvěrný přítel J. Neruda, že domnělý podivuhodný básník Barák je jakýsi radioaktivní isotop závratné tvůrčí potence Nerudovy“ (Králík (ed.) 1958: 74–75).

Králíkova hypotéza se setkala se silnou kritikou. V rovině literárněhistorické argumentace stojí za připomenutí zejm. reakce Felixe Vodičky (1958) a Emanuela Macka (1974). Z našeho pohledu je ovšem nejpodstatnější reakce Pavla Vašáka opírající se o tehdejší stylometrické metody.

4.2.1 Atribuce provedená Pavlem Vašákem

Vašákova atribuce povídky *Kříž pod Petřínem* shrnutá v knize *Metody určování autorství* (1980),[48] resp. zamítnutí hypotézy, že jejím autorem je Jan Neruda, vychází ze srovnání daného textu se sedmnácti Nerudovými povídkami publikovanými mezi lety 1858–1860, konkrétně ze srovnání

(1) průměrné délky vět měřené počtem slov, a to vět různě definovaných a tříděných: věta obecně, věta v řeči vypravěče, věta v řeči postav, uvozovací věta aj. (celkem 10 různých typů);

(2) vzájemného poměru počtu vět zakončených substantivem a vět, jejichž předposledním slovem je substantivum;

(3) průměrné délky slova měřené počtem písmen;

(4) vzájemného poměru slov o čtyřech a pěti písmenech.

U jednotlivých charakteristik Vašák ukazuje, že text *Kříže pod Petřínem* stojí obvykle mimo interval vymezený minimy a maximy zjištěnými u Nerudových textů, případně na jeho okraji, z čehož dovozuje, že „autorem je nonNeruda, tj. je jím v logice důkazu sporem skutečně Josef Barák“ (Vašák 1980: 97).

Vašákova metoda je tedy příkladem jednoduché jednorozměrné statistické analýzy. Přestože pracuje s vícerozměrnými daty (každý text je charakterizován sadou 13

48 Navazuje na starší články k autorství textů připisovaných Josefu Barákovi: Vašák: 1972, 1974.

číselných údajů), analyzuje se vždy jen rozdělení jednoho z nich (průměrná délka věty v řeči vypravěče *Kříže* je kratší než v Nerudových povídkách ⇒ autorem není Neruda; průměrná délka slova je kratší než v Nerudových povídkách ⇒ autorem není Neruda) a nikoliv informace vyplývající z celé sady. Podívejme se proto nejprve na některá Vašákova data se zřetelem k tomu, co z nich lze vyčíst jako z celku.

Tabulka 4.5 (dle Vašák 1980: 190) uvádí pro text *Kříže pod Petřínem* (K) a 15 Nerudových povídek (N1–N17)[49] průměrnou délku věty v řeči vypravěče (X), věty v řeči postav (P), uvozovací věty (Y) a průměrnou délku tří předešlých typů dohromady (X + P + Y).

	X	**P**	**Y**	**X + P + Y**
N1	11,83	7,77	7,27	11,26
N2	13,96	9,33	8,74	10,52
N3	13,17	10,98	7,67	11,57
N4	16,5	7,64	9,67	14,43
N5	16,37	7,8	12,15	10,34
N6	17,13	6,05	10,75	15,26
N7	16,36	6,65	6,00	11,39
N8	16,17	10,9	8,34	11,82
N9	19,06	13,75	9,46	14,21
N11	16,66	9,14	8,12	10,69
N13	16,22	8,76	10,46	11,81
N14	14,33	6,94	6,48	11,24
N15	14,19	10,27	7,29	11,13
N16	16,11	9,23	3,87	14,91
N17	14,22	9,48	9,46	11,78
K	12,87	5,61	6,14	9,98

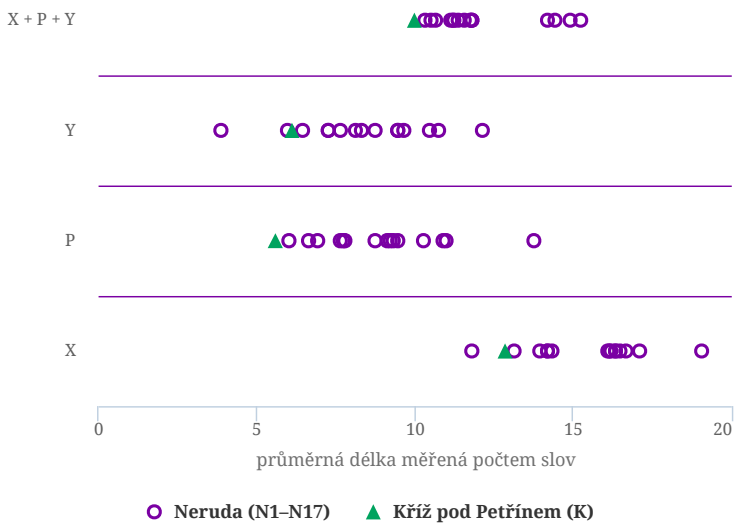
TAB. 4.5: Průměrná délka věty v řeči vypravěče (X), věty v řeči postav (P), uvozovací věty (Y) a průměrná délka tří předešlých typů dohromady (X + P + Y) v textu *Kříže pod Petřínem* (K) a 15 Nerudových povídkách (N1–N15). Měřeno počtem slov.

Tato pozorování slouží Vašákovi jako argument pro zamítnutí Nerudova autorství *Kříže pod Petřínem*: „Vrátíme-li se ke spornému textu [K], zjistíme, že průměrná délka jeho věty typu X + P + Y opět leží zcela na okraji textů Nerudových ze sledovaného období, neboť příslušný průměr je pouze 9,98. Tuto skutečnost lze opět považovat za odmítnutí hypotézy o Nerudově autorství Kříže pod Petřínem. Shrneme-li rozbor vět typu *X* (řeč autora), *P* (řeč postav), *Y* (uvozovací věta), *X* + *P* + *Y* (věta vyplývající z členění na pásmo vypravěče a postav), *T* – *T* (formálně vzatá věta tečka–tečka) z hlediska atribuce textu Kříž pod Petřínem, můžeme konstatovat, že sporný text se vždy situoval na okraji textů Nerudových, respektive se přímo vymykal z jejich řady“ (Vašák 1980: 78).

Vašákovu interpretaci je ale třeba přinejmenším korigovat. Pro jasnější představu se podívejme na grafické vyjádření hodnot z TAB. 4.5 (OBR. 4.11).

Na první pohled je patrné, že v žádné ze čtyř kategorií nemáme co do činění se situací, která by umožňovala výše uvedené závěry, tedy s jasně ohraničeným rozdělením hodnot v Nerudových povídkách (invariant) a vymykající se hodnotou v *Kříží pod Petřínem*:

49 Zachováváme původní Vašákovu notaci s jedinou výjimkou – *Kříž pod Petřínem* značíme K namísto X (bylo poněkud zmatečně užíváno pro označení dvou různých kategorií). Soupis titulů analyzovaných Nerudových povídek viz Vašák 1980: 66. Nepracujeme zde s povídkami N10 (*Kassandra*) a N12 (*Z tobolky redaktorovy*), kde nejsou některé typy analyzovaných vět doloženy.



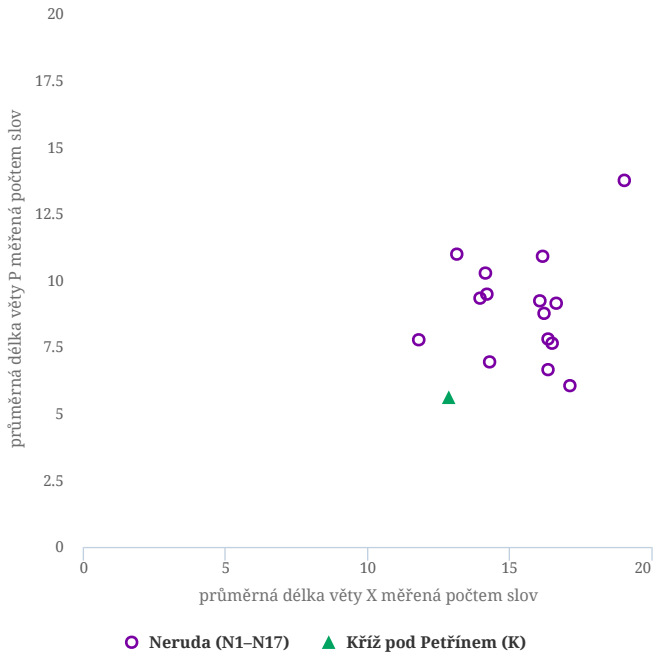
OBR. 4.11: Průměrná délka věty v řeči vypravěče (X), věty v řeči postav (P), uvozovací věty (Y) a průměrná délka tří předešlých typů dohromady (X + P + Y) v textu *Kříže pod Petřínem* a 15 Nerudových povídkách. Měřeno počtem slov.

(1) Ani v jednom případě nevykazuje *Kříž* hodnotu nijak výrazně odlehlou od hodnot zjištěných u Nerudy („*nevymyká se z řady*“), ve dvou případech (věta X, věta Y) nepředstavuje ani hodnotu krajní.

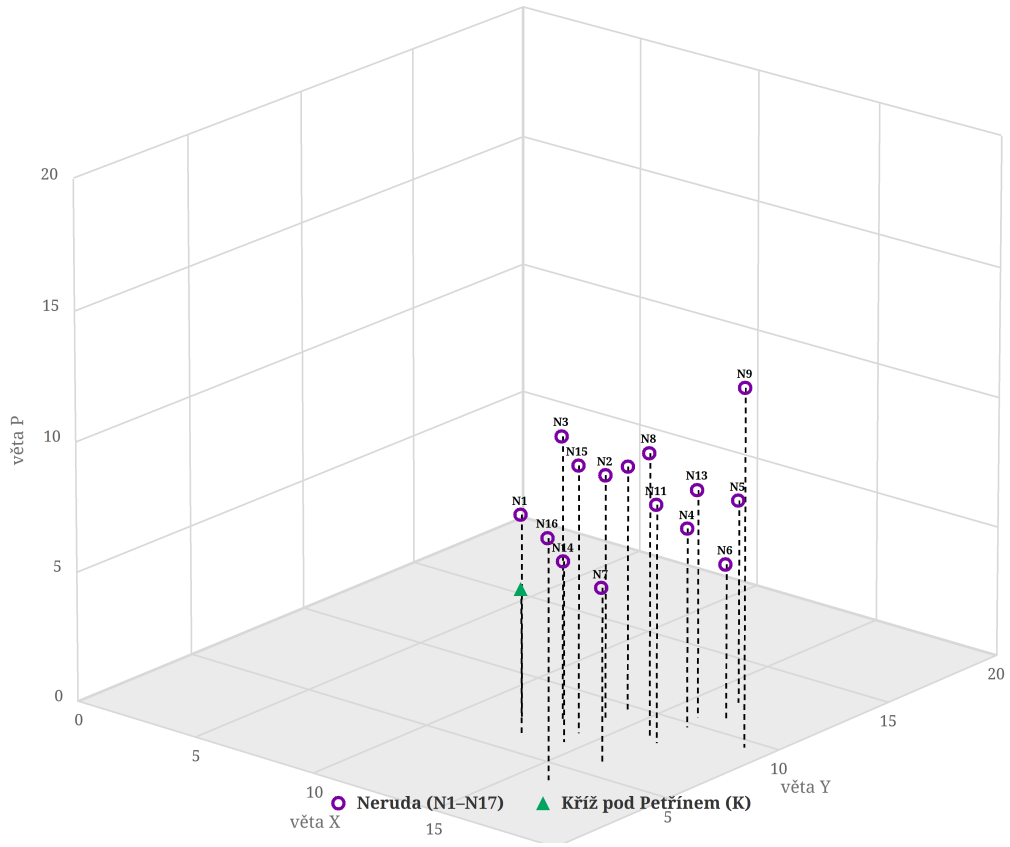
(2) Hodnoty zjištěné v Nerudových textech vykazují poměrně vysokou variabilitu. Ve všech čtyřech kategoriích najdeme mezi nimi i výrazně odlehlé hodnoty.

(3) Zvláštní pozornost si zasluhuje kategorie X + P + Y, kde většina zjištěných hodnot (včetně *Kříže*) spadá do poměrně úzkého intervalu, z něhož se zřetelně vymykají čtyři Nerudovy povídky. Vašák tuto situaci interpretuje paradoxně: „Zdá se, že v tomto případě je již možno mluvit o hledaném invariantu. Tabulka 1 ukazuje, že ze zpracovaných textů se jich dvanáct situuje do velice úzkého intervalu 10,34–11,82! […] Je zajímavé, že čtyři z pěti textů nezapadajících do tohoto intervalu se vyznačují malým relativním zastoupením věty postav (tj. Rozložení P), jmenovitě N4 – 19,69 %, N6 – 9,30 %, N12 – 0,00 %, N16 – 9,50 %“ (Vašák 1980: 77). K tomu dodejme, že s relativně nízkým zastoupením věty typu P (které má vysvětlovat anomálie u Nerudy) se setkáme i v textu *Kříže* (= 19,97; srov. Vašák 1980: 191), jehož vzdálenost od průměru hodnot ležících v intervalu ⟨10,34; 11,82⟩ je mnohem menší než u čtyř výše uvedených textů.

Podívejme se ještě na zobrazení hodnot zjištěných u prvních dvou typů vět (X, P) v dvourozměrném grafu (OBR. 4.12) a následně na zobrazení hodnot zjištěných u prvních tří typů vět (X, P, Y) v trojrozměrném grafu (OBR. 4.13). Pokud by byly Vašákovy závěry správné, mohli bychom očekávat, že čím více dimenzí budeme přidávat, tím zřetelněji se budou Nerudovy texty hromadit do jednoho shluku, z něhož se bude text *Kříže* čím dál tím víc vymykat. Je ale patrné, že k ničemu takovému nedochází. *Kříž* se sice umísťuje na okraji shluku, nijak zvlášť z něj ale nevybočuje. Zřetelně se naopak od ostatních textů odlišuje Nerudova povídka N9 (*Za půl hodiny*).



OBR. 4.12: Průměrná délka věty v řeči vypravěče (X) a průměrná délka věty v řeči postav (P) v textu *Kříže pod Petřínem* a 15 Nerudových povídkách. Měřeno počtem slov.



OBR. 4.13: Průměrná délka věty v řeči vypravěče (X), věty v řeči postav (P) a uvozovací věty (Y) v textu *Kříže pod Petřínem* a 15 Nerudových povídkách. Měřeno počtem slov.

Vašákův model lze tedy označit za značně nespolehlivý. Text, který je brán jako prokazatelně Nerudův (N9: *Za půl hodiny*), by chybně klasifikoval jako text jiného autora, při uplatnění původních vágních kritérií „situuje se na okraji nebo se vymyká“ bychom pak mohli chybně klasifikovat i většinu ostatních Nerudových textů (např. N3: *Erotomanije*, N6: *Mému vrabci* aj.).

K tomu dále dodejme:

(1) Tři Vašákem analyzované charakteristiky (průměrná délka věty, průměrná délka slova a zastoupení slov různé délky) byly při rozsáhlém empirickém testování na anglicky psaném materiálu vyhodnoceny jako vůbec nejslabší, nejméně spolehlivé ukazatele autorství (Grieve 2007).

(2) Vašákovy závěry vycházejí v několika případech z velice malých vzorků. Nejzásadnější je to u analýzy průměrné délky uvozovací věty. V *Kříži pod Petřínem* je průměr vypočítán z pouhých 7 (!) v něm nalezených vět tohoto typu, u jednotlivých Nerudových povídek je to v průměru 18,9 vět na povídku, ve třech případech (*Erotomanije, Byl darebákem, Pražské pověsti*) pak méně než 10 (srov. tabulka 2, Vašák 1980: 191).[50]

(3) Jednorozměrné modely pro atribuci autorství vykazují obecně velice nízkou úspěšnost (srov. Juola 2006; Koppel–Schler–Argamon 2009), a z toho důvodu se v dnešní době prakticky neužívají.

Otázku autorství textů připisovaných Josefu Barákovi jsme se proto rozhodli znovu otevřít, tentokrát analýzou objemnější veršované části díla

4.2.2 Data

Pro naše potřeby by bylo samozřejmě nejvhodnější porovnávat básně podepsané Josefem Barákem (dále „sporné básně“) s korpusem obsahujícím básnická díla pouze dvou autorů: Josefa Baráka a Jana Nerudy. Žádné jiné Barákovy básně krom těch, jejichž autorství bylo zpochybněno, ovšem k dispozici nemáme.

Sporné básně jsme proto zkusili porovnat z hlediska versologických i běžně užívaných stylometrických rysů s trénovacím korpusem obsahujícím vzorky pocházející ze sbírek časově blízkých předpokládané době vzniku Barákových básní. Je zřejmé, že taková analýza nemůže přímo sloužit jako argument pro Barákovo ne-autorství / Nerudovo autorství sporných básní, umožní nám ale učinit závěry o stylové podobnosti sporných básní a básní Jana Nerudy ve srovnání s dobovým kontextem.

Vzhledem k tomu, že sporné básně jsou co do metrického repertoáru poměrně různorodé, rozhodli jsme se versologický model postavit pouze na charakteristikách rýmu a četnostech hlásek. Do každého vzorku jsme tak zahrnuli 250 veršů bez ohledu na jejich rozměr, pouze s podmínkou, aby se jednalo o verše s ženskou klauzulí a vzorek obsahoval přinejmenším 50 rýmových párů. Žádná báseň nebyla rozdělena mezi více vzorků. Tímto způsobem byl vytvořen jeden vzorek pro sporné básně a 54 vzorků pro trénovací korpus (TAB. 4.6).

50 Vašák sice sám uvádí, že „situace [je] komplikována malým počtem Nerudových uvozovacích vět v jednotlivých povídkách a získané průměry je nutno brát s rezervou“ (1980: 77), později ale tuto charakteristiku bere bez výhrad za jeden z důkazů Barákova autorství (viz Vašák 1980: 78, 90).

Autor	Básnické sbírky	# vzorků
Heyduk, Adolf	*Básně 1* (1859); *Básně 2/1* (1864)	12
Hálek, Vítězslav	*Alfréd* (1858); *Večerní písně* (1859); *Mejrima a Husejn* (1859)	8
Jahn, Jiljí Vratislav	*Růženec* (1863)	4
Martinec, Jaroslav	*Mladému pokolení* (1863)	3
Neruda, Jan	*Hřbitovní kvítí* (1858); *Knihy veršů* (1868)	10
Šolc, Václav	*Prvosenky* (1868)	5
Pfleger, Moravský Gustav	*Dumky* (1857); *Duma* (1858)	12

TAB. 4.6: Složení trénovacího korpusu.

4.2.3 Míry z rodiny Delta

Vzhledem k nízkému počtu vzorků u některých autorů lze předpokládat, že výše užívanému Support Vector Machine (SVM) budou v tomto případě co do úspěšnosti konkurovat míry z rodiny Delta (kap. 1.3.1, 1.3.2). Vedle SVM jsme proto testovali i Burrowsovu Deltu (Δ), Argamonovu kvadratickou Deltu (Δ_Q) a Smith-Aldridgovu kosinovou Deltu ($\Delta_\angle$).

V prvním kroku byla provedena validace versologického modelu na trénovacím korpusu. Nejvyšší úspěšnosti dosáhl oproti očekávání SVM (0,7963), dále $\Delta_\angle$ (0,7407), znatelně nižší pak Δ (0,5556) a Δ_Q (0.4074). I tyto hodnoty ale více než dvojnásobně převyšují vypočtenou hodnotu *random baseline* (0,1721).

Dále byla provedena validace modelů založených na:

(1) relativních četnostech n nejfrekventovanějších slovních tvarů,

(2) relativních četnostech n nejfrekventovanějších lemmat,

(3) relativních četnostech n nejfrekventovanějších znakových bigramů

(4) relativních četnostech n nejfrekventovanějších znakových trigramů

(5) relativních četnostech n nejfrekventovanějších znakových tetragramů,

a to pro 20 různých nastavení: $n \in \{50, 100, 150, ..., 1000\}$. Celkem tak bylo vytvořeno 5 × 20 = 100 modelů. U každého z nich byl pak navíc validován ještě model kombinující příslušné rysy s rysy versologickými. Validace byla opět provedena se čtyřmi klasifikátory: Δ, Δ_Q, $\Delta_\angle$ a SVM. Výsledky zobrazuje OBR. 4.14.

Je patrné, že:

(1) Ve všech případech je nejúspěšnějším klasifikátorem $\Delta_\angle$. Nadále se proto budeme zabývat pouze tímto klasifikátorem.

(2) U $\Delta_\angle$ dosahuje kombinovaný model téměř vždy lepších výsledků než model bez versologických rysů.

(3) U $\Delta_\angle$ (stejně jako u ostatních měr z rodiny Delta) narůstá u nekombinovaných modelů úspěšnost se zvyšující se hodnotou m až zhruba k hranici n = 500, po níž zůstává úspěšnost poměrně stabilní (výjimku představují znakové trigramy, kde se stoupající trend neprojevuje). Zapojení versologických rysů tento trend do značné míry neutralizuje.

(4) Z testovaných jednotek dosahují u $\Delta_\angle$ nejlepších výsledků lemmata (s výjimkou $n \in \{50, 100\}$ se úspěšnost nekombinovaného modelu pohybuje nad hranicí 0,85, úspěšnost kombinovaného modelu nad hranicí 0,9).

OBR. 4.14: Úspěšnost klasifikátorů $\Delta_\angle$, Δ, Δ_Q a SVM při použití různých rysů (n nejfrekventovanějších slovních tvarů, lemmat a znakových 2–4gramů) a různých hladinách n (50, 100, 150, …, 1000). Přerušovaná čára udává úspěšnost těchto modelů, plná čára udává úspěšnost modelu, v němž jsou dané rysy zkombinovány s rysy versologickými.

Teď se podívejme, které vzorky obsažené v trénovacím korpusu jsou v jednotlivých modelech založených na $\Delta_\angle$ vyhodnoceny jako nejbližší soused sporných básní. U kombinovaných modelů založených na lemmatech, slovních tvarech, znakových bigramech a znakových trigramech byl ve všech případech za stylisticky nejpodobnější některý ze vzorků pocházejících z Nerudových *Knih veršů*, ve většině případů Nerudův vzorek č. 3 obsahující verše z básní *Kolovrátek* a *O Šimonu Lomnickém*. Pouze u znakových bigramů,

a to navíc u nízkých hladin n, byl za nejbližšího souseda označen vzorek Adolfa Heyduka č. 11 ($n \in \{100, 150, 200\}$. Nerudovi tedy přisoudilo sporné básně 97 % z těchto modelů. Dodejme, že z nekombinovaných modelů označuje jeden ze vzorků z *Knih veršů* za nejbližšího souseda 91 % modelů a za nejbližšího souseda ho označuje i versologický model sám o sobě.

Na základě vysoké úspěšnosti zvolené metody a převahy shodných výsledků můžeme konstatovat, že sporné básně vykazují větší množství společných rysů s ranými Nerudovými básněmi obsaženými ve sbírce *Knihy veršů* než s jakýmikoliv jinými vzorky z trénovacího korpusu. Z toho ale, jak už jsme uvedli výše, nelze vyvozovat, že jejich autorem je skutečně Neruda. Zvolená metoda sice vykazuje vysokou úspěšnost, ale pouze za předpokladu, že *texty skutečného autora jsou v trénovacím korpusu obsaženy*. Jinými slovy, nějaká položka z trénovacího korpusu bude vždy nejbližším sousedem sporných básní, ať už je skutečný autor v trénovacím korpusu přítomen, nebo ne. Výsledek „nelze určit, kdo je autorem sporných básní“ Delta nenabízí. V následující kapitole se podíváme na techniku, která tato omezení umožňuje do určité míry překonat.

4.2.4 Bootstrapovaná kosinová Delta

Boostrapování Delty (Eder 2013) bylo navrženo jako jedna z možných pojistek proti chybné atribuci v případě, kdy skutečný autor není v trénovacím korpusu přítomen: „The procedure [...] displays an accuracy comparable to the state-of-the-art methods used in stylometry, but it is far more sensitive to fake candidates. While the existing methods provide two possible answers to the problem of attribution: *X is the author* or *X is not the author*, the procedue proposed introduces a third answer: *I do not know / I am not sure*, an important safety net against false attribution“ (Eder 2013: 172).

V následujícím textu nejdřív nastíníme princip Ederovy metody (4.2.4.1), poté výsledky, jakých dosáhla na našem materiálu (4.2.4.2).

4.2.4.1 Metoda

Ederova bootstrapovaná Delta vychází z testování robustnosti modelu vůči absenci náhodně vybraných rysů. Při bootstrapování je vytvořen model s dostatečně vysokým počtem analyzovaných rysů. V k iteracích je pak vždy náhodně vybrán náhodně zvolený počet rysů, které jsou z modelu odstraněny. Na takto upraveném modelu je vždy provedeno měření Delta (v našem případě $\Delta_\angle$).

Pro každou dvojici vzorků tak namísto jedné vzdálenosti získáme sadu k vzdáleností. U každé sady je spočten průměr ($\bar{x}$) a 95% interval spolehlivosti, tj. interval $\langle L; U\rangle$ určený 1,96násobkem směrodatné odchylky (σ) nad a pod průměrem:

$$\begin{aligned} L &= \bar{x} - 1{,}96\,\sigma \\ U &= \bar{x} + 1{,}96\,\sigma \end{aligned} \tag{4.3}$$

Při porovnání sporných básní s trénovacím korpusem je uplatněn následující postup:

(1) Z trénovacího korpusu je vybrán vzorek s nejnižší hodnotou průměru vzdáleností od atribuovaného vzorku (t_1) a všechny texty t_2, ..., t_m, jejichž interval spolehlivosti $\langle L_i; U_i\rangle$ se překrývá s intervalem spolehlivosti prvního vzorku $\langle L_1; U_1\rangle$.

(2) Pro každý vzorek t_i je spočteno skóre $c_i \in (0; 1\rangle$, odpovídající míře jistoty, že autor vzorku t_i je autorem sporného textu:

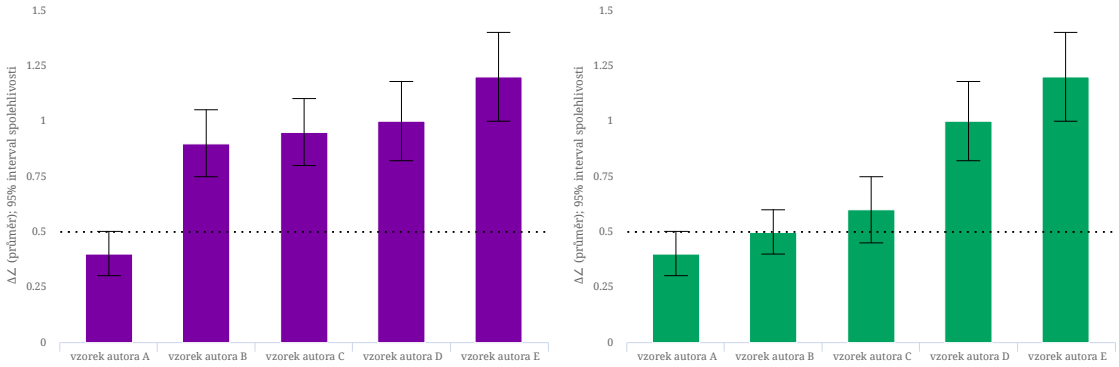
$$c_i = \frac{U_1 - L_i}{\sum_{j=1}^{m} U_1 - L_j} \tag{4.4}$$

Jinými slovy, máme celkové skóre $C = 1$, které je rozděleno mezi vzorky $t_1, t_2, \ldots, t_m$ jako $c_1, c_2, \ldots, c_m$ tak, že

$$\sum_{i=1}^{m} c_i = C = 1 \tag{4.5}$$

(3) Je-li mezi $t_1, \ldots, t_m$ více textů jednoho autora, je výsledné skóre pro každého autora dáno součtem skóre přidělených jeho jednotlivým textům (OBR. 4.15).

Podle Ederových pozorování dochází právě v případech, kdy není skutečný autor v trénovacím korpusu přítomen, k tomu, že je skóre rozděleno mezi větší množství autorů (model není dostatečně robustní).



OBR. 4.15: Příklad: bootstrapovaná Delta. Jednoznačná (vlevo) a nejednoznačná (vpravo) atribuce.

4.2.4.2 Výsledky

Ederovu metodu jsme aplikovali na porovnání sporných básní s trénovacím korpusem a zároveň na porovnání každého vzorku z trénovacího korpusu se souborem tvořeným zbytkem korpusu a spornými básněmi. Jako nejvhodnější soubor rysů jsme zvolili kombinaci 1000 nejfrekventovanějších lemmat a 314 versologických charakteristik. V 10 000 iteracích jsme pak náhodně vybrali $r \in \{0, 1, 2, \ldots, 1000\}$ rysů, které jsme z modelu odstranili, a spočetli hodnoty Δ_Z. Na základě těchto 10 000 měření jsme spočítali výsledné skóre (TAB. 4.7).

Pokud bychom úspěšnost metody posuzovali na základě toho, u kolika vzorků z trénovacího korpusu obdržel nejvyšší skóre jejich skutečný autor, dostali bychom se k 96 % (52 z 54 vzorků). Nejvíce „zmatený“ byl algoritmus v případě Nerudova vzorku č. 4 (obsahujícího verše z básní *Jeník*, *Mrtvá Nevěsta*, *Matka* a *Rubáš* ze sbírky *Knihy veršů*), kde za nepravděpodobnějšího autora označil Adolfa Heyduka, a dále v případě Šolcova vzorku č. 1 (obsahujícího verše od básně *Primula veris* po báseň *Písničkář* ze sbírky *Prvosenky*), kde za stejně pravděpodobného jako skutečného autora označil opět Adolfa Heyduka (Šolc celkově vykazuje poměrně nízkou míru rozpoznatelnosti).

Nota bene, že velice blízká skóre byla přidělena také Nerudovu vzorku č. 3 (obsahujícímu verše z básní *Kolovrátek* a *O Šimonu Lomnickém* ze sbírky *Knihy veršl*) a autorovi sporných básní. Nejpodstatnější ale je, že ve všech případech, kdy bylo jednomu autorovi přiděleno skóre 1 (36 vzorků), jednalo se vždy o atribuci správnou, a že mezi tyto případy spadá i atribuce sporných básní Janu Nerudovi.

Abychom ověřili Ederův předpoklad o citlivosti metody k případům, kdy mezi kandidáty chybí skutečný autor, zopakovali jsme celý proces pro každý vzorek s výjimkou sporných básní ještě jednou. Ze souborů, s nimiž byly vzorky srovnávány, jsme ale tentokrát odstranili vzorky napsané jejich skutečným autorem. Výsledky ukazuje TAB. 4.8.

Už při letmém pohledu si lze všimnout nápadných rozdílů oproti TAB. 4.7. Skóre jsou rozprostřena mezi mnohem větší množství autorů (v TAB. 4.7 vykazuje u jednoho vzorku nenulové skóre v průměru 2,2 autorů, v TAB. 4.8 je to 5,08) a rozdíly mezi nimi jsou znatelně menší (v TAB. 4.7 činí průměrný rozdíl mezi nejpravděpodobnějším a druhým nejpravděpodobnějším autorem v průměru 0,78; TAB. 4.7 je to 0,16). Skóre 1 nebylo v TAB. 4.8 přiděleno žádnému z kandidátů.

Můžeme tak konstatovat, že i metoda, která je do značné míry schopná odhalit případy, kdy mezi kandidáty chybí skutečný autor, označila za nejpravděpodobnějšího autora sporných básní Jana Nerudu.

4.2.5 Shrnutí

Ukázali jsme, že atribuce *Kříže pod Petřínem* provedená Pavlem Vašákem je značně nespolehlivá. Vzhledem k tomu, že Vašákovy práce představovaly dosud jediný stylometrický příspěvek do diskuze o autorství díla připisovaného Josefu Barákovi, otevřeli jsme tuto otázku znovu, tentokrát analýzou veršované části díla. Pomocí Smith–Aldridgovy kosinové Delty jsme ukázali, že sporné básně vykazují velkou míru podobnosti s ranou Nerudovou sbírkou *Knihy veršů* na rovině lexikální (četnost slovních tvarů, četnost lemmat), u jednotek obsahujících informace z různých jazykových rovin (frekvence znakových *n*-gramů) i u versologických rysů. Jako pravděpodobný autor sporných básní byl pak Neruda označen i metodou citlivou k nepřítomnosti skutečného autora v souboru kandidátů (Ederova bootstrapovaná kosinová Delta). Podobnost ovšem mohla být zapříčiněna různými důvody a není na místě zde vyslovovat kategorické soudy o autorství sporných básní.

Vzhledem k absenci dochovaných materiálů a neúplnosti svědectví z doby příprav almanachu *Máj* či redigování *Obrazů života* nemůžeme totiž spolehlivě popsat proces, kterým prošly sporné básně před jejich zveřejněním. *Nota bene*, že sám Neruda v korespondenci zmiňuje, že příspěvky běžně velmi pečlivě redigoval, někdy dokonce „olepšoval“. Nelze proto vyloučit, že zmiňovaná blízkost může být prostě důsledkem toho, že Neruda do původních Barákových textů (minimálně do těch, které prošly jím redigovanými nebo spoluredigovanými periodiky) znatelně zasahoval. Mohli bychom jít dokonce ještě dál a začít uvažovat o možnosti pravidelné přátelské výpomoci v podobě Nerudovy supervize Barákových textů. Vzhledem k absenci rukopisů, které by tuto možnost dokládaly, se ovšem pohybujeme čistě na rovině spekulace.

	autor sporných básní	Heyduk	Hálek	Jahn	Martinec	Neruda	Pfleger	Šolc
Heyduk (1)		**1,00**						
Heyduk (2)		**1,00**						
Heyduk (3)		**1,00**						
Heyduk (4)		0,68	0,05			0,27		
Heyduk (5)		**1,00**						
Heyduk (6)		0,41	0,01	0,16	0,14	0,09		0,18
Heyduk (7)	0,02	0,68	0,03	0,03	0,08	0,09	0,09	
Heyduk (8)	0,03	0,44	0,12		0,06	0,12		0,22
Heyduk (9)	0,02	0,70	0,13		0,06	0,07		0,02
Heyduk (10)		0,87						0,13
Heyduk (11)		**1,00**						
Heyduk (12)		**1,00**						
Hálek (1)			**1,00**					
Hálek (2)			**1,00**					
Hálek (3)			**1,00**					
Hálek (4)			**1,00**					
Hálek (5)			**1,00**					
Hálek (6)			**1,00**					
Hálek (7)			**1,00**					
Hálek (8)			**1,00**					
Jahn (1)				**1,00**				
Jahn (2)				**1,00**				
Jahn (3)				**1,00**				
Jahn (4)				**1,00**				
Martinec (1)				0,04	0,96			
Martinec (2)		0,04		0,25	0,50	0,07		0,14
Martinec (3)		0,03	0,35		0,39	0,21		0,03
Neruda (1)						**1,00**		
Neruda (2)						**1,00**		
Neruda (3)	0,48					0,52		
Neruda (4)	0,15	0,37	0,11			0,30		0,08
Neruda (5)	0,12	0,12	0,08			0,67		
Neruda (6)						**1,00**		
Neruda (7)						**1,00**		
Neruda (8)	0,05	0,11	0,08			0,67	0,01	0,08
Neruda (9)						**1,00**		
Neruda (10)				0,20		0.80		
Pfleger (1)							**1,00**	
Pfleger (2)							**1,00**	
Pfleger (3)							**1,00**	
Pfleger (4)							**1,00**	
Pfleger (5)							**1,00**	
Pfleger (6)							**1,00**	
Pfleger (7)							**1,00**	
Pfleger (8)							**1,00**	
Pfleger (9)							**1,00**	
Pfleger (10)							**1,00**	
Pfleger (11)							**1,00**	
Pfleger (12)							**1,00**	
Šolc (1)		0,42	0,08		0,05	0,03		0,42
Šolc (2)		0,17			0,11		0,11	0,61
Šolc (3)	0,01	0,23	0,04		0,01	0,22	0,06	0,44
Šolc (4)		0,38	0,02	0,02	0,07	0,06		0,44
Šolc (5)		0,07		0,10	0,28		0,12	0,44
sporné básně						**1,00**		

TAB. 4.7: Bootstrapovaná kosinová Delta.

	Heyduk	Hálek	Jahn	Martinec	Neruda	Pfleger	Šolc
Heyduk (1)		0,12	0,12	0,09	0,27	0,09	0,31
Heyduk (2)		0,08	0,07	0,08	0,32	0,11	0,34
Heyduk (3)		0,16	0,07	0,06	0,14	0,28	0,30
Heyduk (4)		0,35			0,65		
Heyduk (5)		0,26	0,02	0,05	0,38	0,09	0,20
Heyduk (6)		0,02	0,25	0,23	0,19		0,32
Heyduk (7)		0,14	0,09	0,16	0,25	0,34	0,02
Heyduk (8)		0,20		0,11	0,32		0,36
Heyduk (9)		0,38		0,13	0,38		0,11
Heyduk (10)					0,42		0,58
Heyduk (11)		0,12		0,06	0,34	0,15	0,33
Heyduk (12)		0,12	0,07	0,10	0,26	0,16	0,29
Hálek (1)	0,28		0,06	0,05	0,25	0,30	0,07
Hálek (2)			0,95	0,01	0,04		
Hálek (3)	0,27			0,18	0,34	0,18	0,03
Hálek (4)	0,01		0,13	0,04	0,67	0,06	0,10
Hálek (5)	0,19		0,05	0,15	0,36		0,25
Hálek (6)	0,34		0,03	0,16	0,20	0,23	0,04
Hálek (7)				0,68	0,32		
Hálek (8)	0,44			0,21	0,24	0,11	
Jahn (1)	0,05	0,18		0,36	0,18	0,10	0,12
Jahn (2)	0,14	0,16		0,18	0,21	0,21	0,09
Jahn (3)	0,16	0,06		0,56		0,13	0,09
Jahn (4)	0,04	0,39		0,34	0,15	0,07	
Martinec (1)	0,25	0,17	0,33		0,09	0,05	0,12
Martinec (2)	0,10		0,46		0,13		0,31
Martinec (3)	0,04	0,55			0,36		0,04
Neruda (1)	0,35	0,31	0,05	0,04		0,01	0,25
Neruda (2)	0,28	0,31	0,08	0,15		0,05	0,12
Neruda (3)	0,34	0,37	0,10	0,12		0,03	0,03
Neruda (4)	0,74	0,15					0,10
Neruda (5)	0,57	0,21	0,02	0,06		0,01	0,13
Neruda (6)	0,32	0,30	0,10	0,11		0,04	0,13
Neruda (7)	0,29	0,29	0,05	0,11		0,17	0,09
Neruda (8)	0,34	0,31	0,04	0,01		0,11	0,19
Neruda (9)	0,28	0,46	0,06	0,13			0,07
Neruda (10)	0,22		0,27	0,52			
Pfleger (1)	0,08	0,15	0,38	0,14	0,09		0,17
Pfleger (2)	0,20	0,21	0,31	0,12	0,04		0,11
Pfleger (3)	0,24	0,20	0,26	0,05	0,07		0,18
Pfleger (4)	0,03	0,48	0,19	0,01	0,15		0,14
Pfleger (5)	0,12	0,56	0,09	0,04	0,05		0,14
Pfleger (6)	0,25	0,30	0,10	0,03	0,13		0,18
Pfleger (7)	0,37	0,05			0,13		0,44
Pfleger (8)	0,38	0,05	0,04	0,04	0,29		0,20
Pfleger (9)	0,31	0,33			0,21		0,15
Pfleger (10)	0,43			0,07	0,09		0,41
Pfleger (11)	0,34	0,22	0,03	0,12	0,19		0,11
Pfleger (12)	0,20	0,32	0,30	0,07	0,01		0,11
Šolc (1)	0,75	0,14		0,07	0,04		
Šolc (2)	0,51	0,02	0,01	0,16	0,05	0,25	
Šolc (3)	0,44	0,10		0,04	0,30	0,13	
Šolc (4)	0,55	0,07	0,05	0,16	0,17		
Šolc (5)	0,14		0,21	0,41	0,01	0,23	

TAB. 4.8: Bootstrapovaná kosinová Delta, není-li mezi kandidáty zahrnut skutečný autor.

Závěr

Cílem této dizertační práce bylo testovat a aplikovat modely rozpoznávání autorství založené na versologických rysech. Shrneme-li nejpodstatnější závěry, které testování přineslo, můžeme konstatovat:

- Úspěšnost rozpoznávání autorství na základě versologických rysů (veršový rytmus, charakteristiky rýmu, četnosti hlásek) převyšuje v české, německé, španělské i anglické versifikaci několikanásobně hodnotu *random baseline*.
- Rozdíly mezi úspěšností rozpoznávání autorství v různých jazycích lze vysvětlovat zejm. jazykově-typologickými faktory (pevnost slovosledu omezující možnosti individualizace rytmu, míra rozvinutosti flexe korelující s velikostí rýmového slovníku).
- Úspěšnost versologických modelů ojediněle převyšuje úspěšnost obvyklých lexikálních modelů.
- Úspěšnost kombinovaných versologicko-lexikálních modelů je zpravidla vyšší než úspěšnost lexikálních modelů samotných.
- Situace, kdy lexikální a versologický model predikuje stejného autora, znamená téměř vždy vyšší spolehlivost atribuce než situace, kdy je predikce provedena pouze lexikálním modelem.

V další části práce byly pomocí versologických modelů, lexikálních modelů a kombinovaných versologicko-lexikálních modelů analyzovány dva soubory textů se sporným autorstvím.

Prvním z nich bylo veršované drama *The Famous History of the Life of King Henry the Eight* poprvé otištěné v Prvním foliu her Williama Shakespeara. Výsledky kombinovaného versologicko-lexikálního modelu ukázaly, že je vysoce pravděpodobné, že se na textu hry kromě Williama Shakespeara autorsky podílel i John Fletcher, zatímco autorská účast Philipa Massingera je značně nepravděpodobná. Dále bylo ukázáno, že technika CUSUM, kterou Thomas Merriam využil k atribuci jednotlivých veršů, není příliš spolehlivá a nese s sebou vysoké riziko falešně pozitivních výsledků. Klasifikace krátkých úseků pomocí techniky klouzavé atribuce (*rolling attribution*) založená na kombinaci versologických a lexikálních rysů se oproti tomu ukázala jako vysoce spolehlivá (úspěšnost na trénovacích datech ~0,999). Při aplikaci na text *H8* se do značné míry potvrdila hypotéza o rozdělení autorství mezi Shakespeara a Fletchera formulována v roce 1850 Jamesem Speddingem a Samuelem Hicksonem.

Druhým případem byly básně publikované pod jménem Josefa Baráka, u nichž byla formulována hypotéza, že jejich skutečným autorem je Jan Neruda. Ukázali jsme, že atribuce prozaické části díla publikovaného pod Barákovým jménem, provedená Pavlem Vašákem a popírající Nerudovo autorství, je značně nespolehlivá. Ukázali jsme, že kombinace versologických a lexikálních rysů funguje při použití měr z rodiny Delta, (zejm.

varianty založené na kosinové podobnosti a její bootstrapované variantě), funguje u poezie 50. let 19. století jako poměrně spolehlivý indikátor autorství. Na tomto základě jsme pak konstatovali vysokou míru stylistické podobnosti mezi analyzovanými básněmi a Nerudovou poezií pocházející zhruba z období, kdy byly publikovány. Jako možné vysvětlení jsme formulovali hypotézu (kterou se zatím nepodařilo na archivních materiálech ověřit), že Neruda do rukopisů zaslaných do redakce Barákem znatelně zasahoval, případně svému příteli s psaním „pomáhal“.

Má-li mít tato práce nějaký zcela jednoznačný závěr, pak ten, že versologické rysy představují relevantní stylometrický indikátor a (přestože jejich extrahování z textu je obvykle znatelně komplikovanější než u tradičně využívaných rysů, jako jsou četnosti slovních tvarů, lemmat, znakových či slovních *n*-gramů) mohou a měly by být při atribuci autorství básnických textů využívány.

Literatura

Abney, S. (2007). Semisupervised Learning for Computational Linguistics. Boca Raton – London – New York: CRC.

Al-Falahi, A. – Ramdani, M. – Bellafkih, M. (2017). Machine learning for authorship attribution in Arabic poetry. International Journal of Future Computer and Communication 6(2), 42–46.

Alexander, P. (1931). Conjectural History, or Shakespeare‘s Henry VIII. Essays and Studies 16, 85–119.

Altmann, G. (1966a). The measurement of Euphony. In J. Levý (ed.), Teorie verše I. Sborník brněnské versologické konference, 13.–16. května 1964. Brno: UJEP, 263–264.

Altmann, G. (1966b). Binomial Index of Euphony for Indonesian Poetry. Asian and African Studies 2, 62–67.

Antilla, A. – Heuser, R. (2016). Phonological and metrical variation across genres. Proceedings of the Annual Meeting on Phonology 3. Linguistic Society of America: Washington D.C.

Argamon, S. (2008). Interpreting Burrows‘s Delta: Geometric and probabilistic foundations. Literary and Linguistic Computing 23(2), 131–147.

Baayen, R. H. (2001). Word Frequency Distributions. Dordrecht: Kluwer Academic Publishers.

Bělyj, A. (1910). Symbolizm: Kniga Statej. Moskva: Musaget.

Bobenhausen, K. (2011). The Metricalizer. Automated metrical markup for German poetry. In C. Küper (ed.), Curent Trends in Metrical Analysis. Frankfurt am Main et al.: Peter Lang, 119–131.

Bobenhausen, K. – Hammerich, B. (2015). Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme Metricalizer². Langages 199, 67–87.

Boyle, R. (1886). Herny VIII. An investigation into the origin and authorship of the play. Transactions of the New Shakspere Society 1880–86, 443–487.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167.

Burrows, J. F. (1989). »An ocean where each kind«: Statistical analysis and some major determinants of literary style. Computer and the humanities 23(4–5), 309–321.

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

Burrows, J. F. (2003). Questions of authorship: attribution and beyond. Computers and the Humanities 37(1), 5–32.

Burrows, J. F. – Hassall, A. J. (1988). Anna Boleyn and the Authenticity of Fielding's Feminine Narrative. Eighteenth Century Studies, 21(4), 427–453.

Craig, H. – Kinney, A. F. (2009). Shakespeare, Computers and the Mystery of Authorship. Cambridge: Cambridge University Press.

Čech, R. – Popescu, I.-I. – Altmann, G. (2011). Euphony in Slovak lyric poetry. Glottometrics 22, 5–16.

Čech, R. – Popescu, I.-I. – Altmann, G. (2014). Metody kvantitativní analýzy (nejen) básnických textů. Olomouc: Univerzita Palackého.

Červenka, M. (1971). Osmislabičná řada ve verši a v próze. In idem, Statistické obrazy verše. Praha: ÚČSL AV ČR, 31 – 50.

Červenka, M. (1998). Máchovo místo ve vývoji českého verše. Česká literatura 46(5), 485–489.

Červenka, M. (2002). Hlásková instrumentace. In M. Červenka, M. Jankovič, M. Kubínová, M. Langerová (eds.), Pohledy zblízka. Zvuk, význam, obraz. Praha: Torst, 7–54.

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

Červenka, M. – Sgallová, K. (1978). Český verš. In Z. Kopczyńska, L. Pszczołowska (eds.), Słowiańska metryka porównawcza I. Słownik rytmiczny i sposoby jego wykorzystania. Wrocław et al.: Ossolineum, 45–94.

Diederich, J. – Kindermann, J. – Leopold, E. – Paass, G. (2003). Authorship attribution with support vector machines, Applied Intelligence 19(1), 109–123.

Dobritsyn, A. (2016). Rhythmic entropy as a measure of rhythmic diversity (The example of Russian iambic tetrameter). Studia Metrica et Poetica 3(1), 33–52.

Eddy, H. (1887). The characteristic curve of composition. Science 9(216), 297.

Eder, M. (2011). Style markers in authorship attribution. A cross-language study of the authorial fingerprint. Studies in Polish Linguistics 6, 99–114.

Eder, M. (2013). Bootstrapping Delta: A safety net in open-set authorship attribution. Digital Humanities 2013: Conference Abstracts. Lincoln: University of Nebraska-Lincoln, 169–172.

Eder, M. (2015). Does size matter? Authorship attribution, short samples, big problem. Digital Scholarship in the Humanities 30(2), 167–182.

Eder, M. (2016). Rolling stylometry. Digital Scholarship in the Humanities 31(3), 457–469.

Eder, M. (2017). Short samples in authorship attribution: A new approach. Digital Humanities 2017: Conference abstracts. Montreal: McGill University, 221–224.

Ege, K. (1922). Shakespeare‘s Anteil an „Henry VIII“. Shakespeare Jahrbuch 58, 99–115.

Eisen, M. – Riberio, A. – Segarra, S. – Egan, G. (2017). Stylometric analysis of early modern period English plays. Digital Scholarship in the Humanities 33(3), 500–528.

Farnham, W. (1916). Colloquial contractions in Beaumont, Fletcher, Massinger and Shakespeare as a test of authorship. Publications of the Modern Language Association of America 31, 326–358.

Fleay, F. G. (1874a). On the authorship of The Taming of the Shrew. Transactions of the New Shakspere Society 1, 85–129.

Fleay, F. G. (1874b). On the authorship of Timon of Athens. Transactions of the New Shakspere Society 1, 130–194.

Fleay, F. G. (1874c). A fresh confirmation of Mr. Spedding‘s division and date of the play of Henry VIII. Transactions of the New Shakspere Society 1, appendix 23.

Fleay, F. G. (1874d). Mr. Hickson‘s division of The Two Noble Kinsmen, confirmed by Metrical Tests. Transactions of the New Shakspere Society 1, appendix 61–64.

Fleay, F. G. (1876). Shakespeare Manual. London: Macmillan.

Fleay, F. G. (1885). Mr. Boyle‘s theory as to „Henry VIII“. Athenæum 2994, 14. 3., 355.

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

Forstall, C. W. – Jacobson, S. L. – Scheirer, W. J. (2011). Evidence of intertextuality: investigating Paul the Deacon‘s Angustae Vitae. Literary and Linguistic Computing 26(3), 285–296.

Forstall, C. W. – Scheirer, W. J. (2010). A statistical stylistic study of Latin elegiac couplets. 2010 Chicago Colloquium on Digital Humanities and Computer Science Abstracts, Evanston: Northwestern University, 21–22. 11.

Fucks, W. (1952). On mathematical analysis of style. Biometrika 39(1–2): 122–129.

Furnivall, F. J. (1874a). The founder‘s prospectus of the New Shakspere Society. Transactions of the New Shakspere Society 1, appendix 6–7.

Furnivall, F. J. (1874b). Another fresh confirmation of Mr. Spedding‘s division and date of the play of Henry VIII. Transactions of the New Shakspere Society 1, appendix 24.

Furnivall, F. J. (1874c). Mr. Hickson‘s division of The Two Noble Kinsmen, confirmed by the the stopt-line test. Transactions of the New Shakspere Society 1, appendix 64–65.

Gebauer, J. (1886). Potřeba dalších zkoušek rukopisu Královedvorského a Zelenohorského. Athenaeum 3, 152–168.

Giesbrecht, E. – Evert, S. (2009). Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In I. Alegria – I. Leturia – S. Sharoff (eds.), Proceedings of the 5th Web as Corpus Workshop (WAC5). San Sebastian.

Göhring, A. (2009). Spanish Expansion of a Parallel Treebank. Magisterská diplomová práce. Universität Zürich.

Grieve, J. (2005). Quantitative authorship attribution: A history and an evaluation of techniques. Master thesis. Simon Fraser University.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. Literary and Linguistic Computing 22(3), 251–270.

Grzybek, P. (2014). The emergence of stylometry: prolegomena to the history of term and concept. In: Kroó, Katalin; Torop, Peeter (eds.), Text within Text – Culture within Culture. Budapest, Tartu: L‘Harmattan, 58–75.

Hart, A. (1941). Vocabularies of Shakespeare‘s Plays. The Review of English Studies 19(74), 128–140.

Hearst, M. A. (1998). Support vector machines. IEEE Intelligent Systems 13(4), 18–28.

Herdan, G. (1956). Chaucer‘s authorship of the Equatorie of the Planetis: The use of romance vocabulary as evidence. Language 32(2), 254–259.

Hickson, S. (1850). Who wrote Shakespeare‘s Henry VIII. Notes and Queries 2, 198.

Holmes, D. I. (1994). Authorship attribution. Computers and the Humanities 28(2), 87–106.

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

Hoover, D. L. (2003). Another perspective on vocabulary richness, Computers and the Humanities 37(2), 151–178.

Hoover, D. L. (2004a). Testing Burrows‘s Delta. Literary and Linguistic Computing 19(4), 453–475.

Hoover, D. L. (2004b). Delta Prime? Literary and Linguistic Computing 19(4), 477–495.

Hope, J. (1994/2009). The Authorship of Shakespeare‘s Plays. Cambridge: Cambridge University Press.

Horsmann, T. – Erbs, N. – Zesch, T. (2015). Fast or accurate? A comparative evaluation of PoS tagging models. In Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL-2015)

Horton, T. B. (1987). The Effectiveness of the Stylometry of Function Words in Discriminating between Shakespeare and Fletcher. Dizertační práce. University of Edinburgh.

Hoy, C. (1957). The Shares of Fletcher and his collaborators in the Beaumont and Fletcher Canon II. Studies in Bibliography 9, 143–162.

Hoy, C. (1962). The shares of Fletcher and his collaborators in the Beaumont and Fletcher Canon VII. Studies in Bibliography 15, 71–90.

Ibrahim, R. – Plecháč, P. – Říha, J. (2013). Úvod do teorie verše. Praha: Akropolis

Ingram, J. K. (1874). On the „weak endings“ of Shakspere, with some account of the history of the verse tests in general. Transactions of the New Shakspere Society 1, 442–456.

Jackson, M. P. (1997). Phrase lengths in Henry VIII. Shakespeare and Fletcher. Notes and Queries 44(1), 75–80.

Jackson, M. P. (2013). All Is True? Or Henry VIII: Authors and ideologies. Notes and Queries 60(3), 441–444.

Jacobs, A. M. (2018). The Gutenberg English Poetry Corpus: Exemplary quantitative narrative analyses. Frontiers in Digital Humanities 5:5. doi: 10.3389/fdigh.2018.00005

Jakobson, R. (1935). K časovým otázkám nauky o českém verši. Slovo a slovesnost 1(1), 46–53.

Jakobson, R. (1938/1995). K popisu Máchova verše. In idem: Poetická funkce. Jinočany: H & H, 427–476

Jannidis, F. – Pielström, S. – Schöch, C. – Vitt, T. (2015). Improving Burrows’ Delta. An empirical evaluation of text distance measures. Digital Humanities Conference 2015, Sydney.

Jirát, V. (1931–1932). Hudebnost Máchova rýmu. Časopis pro moderní filologii 18(1–2), 24–34 & 147–157.

Juola, P. (2006). Authorship attribution. Foundations and Trends in Informational Retrieval 1(3), 233–334.

Kestemont, M. (2014). Function words in authorship attribution. From black magic to

theory? Proceedings of the3rd Workshop on Computational Linguistics for Literature. Göteborg: ACL, 59–66.

Kestemont, M. – Haverals, W. (2018). Metrical analyses of medieval Dutch poetry for the purpose of genre and authorship analysis [konferenční příspěvek]. Plotting Poetry II: Bringing Deep Learning to Computational Poetry Analysis. Berlin: Freie Universität Berlin, 12.–14. 9. 2018.

Koppel, M. – Schler, J. (2004). Authorship verification as a one-class classification problem. Proceedings of the 21st International Conference on Machine Learning. New York: ACM, 62–68.

Koppel, M. – Schler, J. – Argamon, S. (2009). Computational methods in authorship attribution. Journal of the Association for Information Science and Technology 60(1), 9–26.

Králík, O. (1956). Svědectví Anny Holinové. Host do domu 3(3), 107–109.

Králík, O. (1957). Neruda nebo Barák? Literární noviny 6(47), 6.

Králík, O. (ed.) (1958). Z doby Májů. Olomouc: Krajské nakladatelství v Olomouci.

Kubát, M. (2016). Kvantitativní analýza žánrů. Ostrava: Ostravská univerzita.

Kuboň, V. – Lopatková, M. – Hercig, T. (2016). Searching for a Measure of Word Order Freedom. In B. Brejová (ed.), Proceedings of the 16th ITAT Conference Information Technologies – Applications and Theory. CreateSpace, 11–17.

Langworthy, C. (1931). A verse-sentence analysis of Shakespeare‘s Plays. Publications of the Modern Language Association of America 46(3), 738–751.

Larsen, W. A. – Rencher, A. C. – Layton, T. (1980). Who wrote the Book of Mormon? An analysis of wordprints. BYU Studies Quarterly 20(3), 225–251.

Levý, J. (1962). Izochronie taktů a izosylabismus jako činitelé básnického rytmu. Slovo a slovesnost 23(1–2), 1–8 & 83–94.

Lotman, J. M. – Lotman, M. (1986). Vokrug desjatoj glavy „Evgenija Onegina“. In N. N. Petrunina (ed.), Puškin: Issledovanija i materialy XII. Moskva/Leningrad: Nauka, 124–151.

Macek, E. (1974). Králík kontra Barák. Literární archiv PNP VIII–IX, 495–580.

Malone, Edmond. (1787/1803). A dissertation on parts one, two and three of Henry the Sixth tending to show that those plays were not written originally by Shakespeare. In Plays of William Shakespeare 14. London, 219–263.

Matthews, R. A. J. – Merriam, T. V. N. (1993). Neural computation in stylometry I. An application to the works of Shakespeare and Fletcher. Literary and Linguistic Computing 8(4), 203–209.

Mascol, C. (1888a). Curves of pauline and pseudo-pauline style I. Unitarian Review 30, 452–460.

Mascol, C. (1888b). Curves of pauline and pseudo-pauline style II. Unitarian Review 30, 539–546.

Matthews, R. – Merriam, T. (1993). Neural computation in stylometry I; An application to the works of Shakespeare and Fletcher. Literary and Linguistic Computing 8(4), 203–209.

Maxwell, B. (1923). Fletcher and Henry the Eighth. The Manly Anniversary Studies in

Language and Literature. Chicago: University of Chicago, 104-112.

Maxwell, B. (1926). Sidelights on Shakespeare; Sidelights on Elizabethan Drama [recenze]. Modern Philology 23(3), 365–372.

Mendenhall, T. C. (1887). The characteristic curves of composition. Science 9(214), 237–249.

Mendenhall, T. C. (1901). A mechanical solution to a literary problem. Popular Science Monthly 9, 97–110.

Merriam, T. (1979). What Shakespeare wrote in ‚Henry VIII‘: part one. The Bard 2(3), 81–94.

Merriam, T. (1980). What Shakespeare wrote in ‚Henry VIII‘: part two. The Bard 2(4), 111–118.

Merriam, T. (2000). Edward III. Literary and Linguistic Computing 15(2), 157–186.

Merriam, T. (2003a). Taylor‘s method applied to Shakespeare and Fletcher. Notes and Queries 50(4), 419–423.

Merriam, T. (2003b). Though this be supplementarity, yet there is method in‘t. Notes and Queries 50(4), 423–426.

Merriam, T. (2005). The Identity of Shakespeare in Henry VIII. Tokyo: Renaissance Institute.

Merriam, T. (2014). A reply to ‚All is True or Henry VIII: Authors and ideologies‘. Notes and Queries 61(2), 253–256.

Merriam T. (2018). Henry VIII, All Is True?. Notes and Queries 65(1), 84–88.

Mikros, G. K. – Perifanos, K. A. (2013). Authorship attribution in Greek tweets using author‘s multilevel *n*-gram profile. In Papers from the 2013 AAAI Spring Symposium. "Analyzing Microtext", 25–27 March 2013. Palo Alto: AAAI Press, 17–23.

Mincoff, M. (1961). Henry VIII and Fletcher. Shakespeare Quarterly 12(3), 239–260.

Mittmann, A. – Pergher, P. H. – dos Santos, A. L. 2019. What rhythmic signature says about poetic corpora. In P. Plecháč – B. Scherr – T. Skulacheva – H. Bermúdez-Sabel – R. Kolár (eds.), Quantitative Approaches to Versification. Praha: Ústav pro českou literaturu AV ČR, 153–172.

de Morgan, A. (1851/1882). To Rev. W. Heald, Aug. 18, 1851 [dopis]. In S. E. de Morgan (ed.), Memoir of Augustus de Morgan. London: Longmans, Green, & co., 214–216.

Mosteller, F. – Wallace D. L. (1964). Inference and Disputed Authorship. Reading: Addison-Wesley.

Mukařovský, J. (1934/2001). Obecné zásady a vývoj novočeského verše. In idem Studie II, Brno: Host, 116–198.

Navarro-Colorado, B. (2015). A computational linguistic approach to Spanish golden age sonnets: metrical and semantic aspects. *Computational Linguistics for Literature NAACL 2015*, Denver (Co).

Navarro-Colorado, B. (2017). A metrical scansion system for fixed-metre Spanish poetry. Digital Scholarship in the Humanities 33(1), 112–127.

Navarro-Colorado, B. – Ribes-Lafoz, M. – Sánchez, N. (2016). Metrical annotation of a large corpus of Spanish sonnets. Representation, scansion and evaluation. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož.

Neidorf, L. – Krieger, M. S. – Yakubek, M. – Chaudhuri, P. – Dexter, J. P. (2019). Large-scale

quantitative profiling of the Old English verse tradition. Nature Human Behaviour 8. 4. 2019, doi: 0.1038/s41562-019-0570-1.

Oliphant, E. H. C. (1891). The Works of Beaumont and Fletcher. Englische Studien 15, 321–360.

Oliphant, E. H. C. (1927). The Plays of Beaumont and Fletcher: An Attempt to Determine their Respective Shares. New Haven.

Oras, A. (1953). »Extra monosyllables« in Henry VIII and the problem of authorship. Journal of English and Germanic Philology 52, 198–213.

Partridge, A. C. (1949). The Problem of Henry VIII Reopened. Some Linguistic Criteria for the Two Styles Apparent in the Play. Cambridge: Bowes & Bowes.

Pelillo, M. (2014). Alhazen and the nearest neighbor rule. Pattern Recognition Letters 38, 34–37.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers 10(3), 61–74.

Plecháč, P. (2016). Czech verse processing system KVĚTA: Phonetic and metrical components. Glottotheory 7, 159–174.

Plecháč, P. (2018). A collocation-driven method of discovering rhymes (in Czech, English, and French poetry). In M. Fidler – V. Cvrček (eds.), Taming the Corpus. From Inflection and Lexis to Interpretation. Cham: Springer, 79–95.

Plecháč, P. – Bobenhausen, K. – Hammerich, B. (2018). Versification and authorship attribution. Pilot study on Czech, German, Spanish, and English Poetry. Studia Metrica et Poetica 5(2), 29–54

Plecháč, P. – Flaišman, J. (2017). Problém Barák–Neruda z pohledu současné stylometrie. Česká literatura 65(5), 743–769.

Plecháč, P. – Kolár, R. (2015). The Corpus of Czech Verse. Studia Metrica et Poetica 2(1), 107–118.

Popescu, I.-I. – Altmann, G. – Grzybek, P. – Jayaram, B. D. – Köhler, R. – Krupa, V. – Mačutek, J. – Pustet, R. – Uhlířová, L. – Vidya, M. N. (2009). Word Frequency Studies. Berlin: De Gruyter.

Popescu, I.-I. – Lupea, M. – Tatar, D. – Altmann, G. (2015). Quantitative Analysis of Poetic Texts. Berlin: De Gruyter.

Rybicki, J – Eder, M. (2011). Deeper Delta across genres and languages: do we really need the most frequent words? Literary and Linguistic Computing 26(3), 315–321.

Rygl, J. (2014). Automatic adaptation of author‘s stylometric features to document types. In P. Sojka et al. (eds.), Text, Speech, and Dialogue. 17Th International Conference. Berlin – Heidelberg: Springer, 59–61.

Rygl, J. (2016). Building Corpora for Stylometric Research. In P. Sojka et al. (eds.), Text, Speech, and Dialogue. 19th International Conference. Berlin – Heidelberg: Springer, 20-27,

Sedlačíková, B. (2012). Historie matematické lingvistiky. Brno: CERM.

Segarra, S. – Eisen, M. – Riberio, A. (2013). Authorship attribution using function words adjacency networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5563–5567.

Seydler, A. (1886). Počet pravděpodobnosti v přítomném sporu. Athenaeum 3, 299–308.

Sherman, L. A. (1888). Some observations upon sentence-length in English prose. The University of Nebraska Studies 1(4), 337–366.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. Proceedings of International Conference on New Methods in Language Processing. Manchester.

Skoumalová, H. (2011). Porovnání úspěšnosti tagování korpusu. In V. Petkevič – A. Rosen (eds.), Korpusová lingvistika Praha 2011/3. Gramatika a značkování korpusů. Praha: Nakladatelství Lidové Noviny, 199–207.

Smith, P. W. H. – Aldridge, W. (2011). Improving authorship attribution: Optimizing Burrows' Delta method. Journal of Quantitative Linguistics 18(1), 63–88.

Spedding, J. (1850). Who wrote Shakespeare‘s Henry VIII. The Gentleman‘s Magazine, 115–123.

Spoustová, D. – Hajič, J. – Votrubec, J. – Krbec, P. – Květoň, P. (2007). The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing. ACL, 67–74.

Stamatatos, E. (2009). A Survey of modern authorship attribution methods. Journal of the Association for Information Science and Technology 60(3), 538–556.

Sykes, H. D. (1919). Sidelights on Shakespeare. Stratford-upon-Avon: The Shakespeare Head Press.

Šapir, M. (1997). Fenomen Batenkova i problem mistifikacii (lingvistichovedčeskij aspekt 1–2). Philologica 4, 85–144.

Šapir, M. (1998). Fenomen Batenkova i problem mistifikacii (lingvistichovedčeskij aspekt 3–4). Philologica 5, 49–132.

Tabata, T. (2012). Approaching Dickens‘ style through random forests. Digital Humanities 2012: Conference Abstracts. Hamburg: Universität Hamburg, 388–391.

Tarlinskaja, M. (1987). Shakespeare's Verse: Iambic Pentameter and the Poet's Idiosyncrasies. New York: Peter Lang.

Tarlinskaja, M. (2014). Shakespeare and the versification of English Drama, 1561–1642. Farnham et al.: Ashgate.

Těšitelová, M. – Nebeská, I. – Králík, J. (1976). On the quantitative characteristics of the Czech texts of disputed authorship. Prague Studies in Mathematical Linguistics 5, 119–147.

Thorndike, A. H. (1901). The Influence of Beaumont and Fletcher on Shakespeare. Worcester: Oliver B. Wood.

Tomaševskij, B. V. (1923/2008). Pjatistopnyj jamb Puškina. In Izbrannyje raboty o stiche. Moskva & Sankt Peterburg: Akademia, 140–242.

Vašák, P. (1966). Statistika a sporné autorství. Slovo a slovesnost 27(4), 364–370.

Vašák, P. (1972). Metody ustanovlenija spornogo avtorstva (problema Neruda — Barák). Prague Studies in Mathematical Linguistics III (Praha), s. 143–162.

Vašák, P. (1974). Barákovo autorství jako problém. Česká literatura 22(2), 145–155.

Vašák, P. (1980). Metody určování autorství. Praha: Academia.

Vodička, F. (1958). Ještě jednou: Neruda nebo Barák. Literární noviny VII, 25. 1., 6.

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

Weber, H. (1812). Observations on the participation of Shakespeare in The Two Noble Kinsmen. In idem (ed.), The Works of Beaumont and Fletcher in Fourteen Volumes 13. Edinburgh: J. Ballantyne & Co., 151–169.

Williams, C. B. (1975). Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. Biometrika 62(1), 207–212.

Wimmer, G. – Altmann, G. – Hřebíček, L. – Ondrejovič, S. – Wimmerová, S. (2003). Úvod do analýzy textov. Bratislava: VEDA.

Wolf, F. (1928). Použití počtu pravděpodobnosti k identifikaci textu. Inaugurace rektorů v Brně 1928/29 a 1929/30. Brno: Masarykova Universita, s. 99–105.

Woodward, R. H. – Goldsmith, P. L. (1964). Cumulative sum techniques. Edinburgh: Oliver and Boyd.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. Biometrika 30, 363–390.

Yule, G. U. (1944). The Statistical Study of Literary Vocabulary. Cambridge: CUP.

Zhao, Y. – Zobel, J. (2005). Effective and scalable authorship attribution using function words. In G. G. Lee et al. (eds.), Information Retrieval Technology. AIRS 2005. Lecture Notes in Computer Science. Berlin – Heidelberg: Springer, 174–189.

Zipf, G. K. (1932). Selected Studies on the Principle of Relative Frequency in Language. Cambridge: HUP.

Seznam obrázků

Seznam tabulek

Prohlášení spoluautorů

As a co-author of the article "Versification and authorship attribution. Pilot study on Czech, German, Spanish and English Poetry" (*Studia Metrica et Poetica* 5(2), 2018, 29–54), I declare that sections 2 (*Motivations*) and 3 (*History and Related Works*) were compiled by Petr Plecháč.

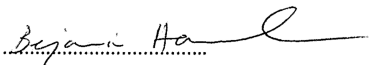
Freiburg im Brsg., 28 June 2019

.....

Klemens Bobenhausen

As a co-author of the article "Versification and authorship attribution. Pilot study on Czech, German, Spanish and English Poetry" (*Studia Metrica et Poetica* 5(2), 2018, 29–54), I declare that sections 2 (*Motivations*) and 3 (*History and Related Works*) were compiled by Petr Plecháč.

Berlin, 28 June 2019



Benjámin Hammerich

Jakožto spoluautor článku „Problém Barák–Neruda z pohledu současné stylometrie“ (*Česká literatura* 65(5), 2017, 743–769) prohlašuji, že můj podíl na výzkumu je představen v oddílu 1 (*Historie problému*) a že autorem kapitol 2 (*Atribuce „Kříže pod Petřínem“ provedená Pavlem Vašákem*), 3 (*Kvadratická kosinová Delta*) a 4 (*Bootstrapovaná kosinová Delta*) je Petr Plecháč.

V Praze 20. 5. 2019

.....
Jiří Flaišman