ROZHLEDY

Problém Barák–Neruda z pohledu současné stylometrie

Petr Plecháč — Jiří Flaišman

Určení autorství (původce) textu patří k jedné z nejsložitějších disciplín v oblasti textové kritiky, potažmo literární vědy obecně. Za autora zkoumaného textu nemůžeme automaticky označit jeho pisatele (podle dochovaného autografu) či toho, kdo daný text publikoval pod svým jménem (díla pseudonymní, autorské krytí, mystifikace). Bezpečné indicie nezajišťují ani další mimotextové skutečnosti, jež provázejí proces zveřejnění textu, jako jsou například doklady o komunikaci s redaktory, nakladatelské smlouvy, různé typy přihlášení se k autorství (včetně např. soudních protokolů),[1] vzpomínky autora nebo jiných pamětníků atd. Jednu z významných cest budovala od počátku moderní filologie, která do služeb atribuce textu vložila celou škálu nástrojů, počínaje stopováním indicií, svědčícím o příslušnosti zkoumaného textu k určitému historickému období, ideovou i stylovou příbuzností se skupinou textů jiného autora, a zejména také zkoumáním jednotlivých jazykových jevů, které prozrazují spojitost s jinou vybranou skupinou textů. U textů moderní literatury jde ve většině případů o snahu identifikovat jednu skupinu textů, u nichž je autorství sporné, s jinou, jejíž autor je bezpečně doložen, a tedy se nesetkáváme s případy — jako je tomu ve starší literatuře —, kdy je zapotřebí alespoň vyloučit, že autorem je osoba, jíž je dílo domněle připisováno.[2]

Vedle nástrojů — řekněme — tradiční atribuce textu, jejichž škálu (výzkum historický, literární, sociologický, filozoficko-náboženský, jazykovědný atd.) v našem prostředí zcela jistě nejlépe reprezentuje tzv. boj o *Rukopisy* z osm-

1 Příkladem může být v protokolech zachycené doznání Petra Bezruče (V. Vaška) během soudního líčení v období první světové války, že je autorem básní *Slezských písní*. Mnozí popírači Vaškova autorství ani toto prohlášení nebrali za bernou minci.

2 Pro tento postup je běžně užíváno termínu *ateteze*.

desátých let 19. století, je v oblasti bohemistických studií poslední čtvrtiny 20. století výzkum v oblasti hledání nových metod atribuce textu spojen zejména se jménem Pavla Vašáka, který se mj. dlouhodobě věnoval dílu K. H. Máchy. Vašák je v našem prostředí také zakladatelem v oblasti aplikace matematických metod pro určování autorství a významným stoupencem exaktních přístupů v textové kritice. Přestože se jím v období sedmdesátých a osmdesátých let propracovávaná metoda dočkala svého shrnutí roku 1993 v příručce *Textologie. Teorie a ediční praxe*,[3] jež měla ve stopách její předchůdkyně sloužit ediční praxi, nutno konstatovat, že nenašla své následovníky a její autor zůstal solitérem.

Výzkum v oblasti sporného autorství textů připisovaných K. H. Máchovi dovedl Pavla Vašáka v sedmdesátých letech k prozkoumání hypotézy, která byla diskutována v předchozích dvou dekádách, a sice zda skutečným autorem textů publikovaných pod jménem Josefa Baráka na přelomu padesátých a šedesátých let 19. století není básník Jan Neruda. V následujícím textu nejprve stručně připomeneme historii sporu o autorství daných textů, dále poukážeme na podstatné slabiny Vašákovy metody a pokusíme se otázku znovu otevřít s využitím metod současné stylometrie.

1. Historie problému

Za významný impulz k budoucí polemice o identitě Barák–Neruda je nutno považovat text přednášky Marie Scherrerové, která připomněla vliv J. Kollára, K. J. Erbena či K. H. Máchy na tvorbu Jana Nerudy, potažmo zdůraznila významné ovlivnění náměty a obrazy, které Neruda ve své rané tvorbě čerpal z díla svých méně známých současníků, publikovaného v dobových sbornících a časopisech. Ve svém příspěvku vyslovila možnost vlivu J. Baráka na Nerudu, když si povšimla podobnosti Barákových veršů s básněmi *Prostých motivů*, neboť díky „námětu melancholického jara […] ladění je stejné a nelze popřít jejich vnější příbuznost“ (SCHERREROVÁ 1949: 114–115). Autorka — žačka Alberta Pražáka — ještě nevyslovila možnost identity obou množin textů, v jejím zorném poli se spíše ocitl Barák jako překvapivý inspirátor již vyzrálé tvůrčí periody v básnickém díle J. Nerudy.

Prvním, kdo se pokusil explicitně přiřknout některý z Barákových beletristických příspěvků Nerudovi, byl v březnu 1956 Oldřich Králík, který věren své tvůrčí vehemenci otevřel tímto krokem vedle sporů o středověké legendy na straně jedné a Máchou a Bezručem na straně druhé další frontu. V březnovém čísle *Hosta do domu* se Králík v komentáři k povídce „Kříž pod Petřínem“, jež jím byla otištěna již s Nerudovým jménem, snažil rehabilitovat svědectví A. Holinové, kterou chápe jako hodnověrnou svědkyni událostí kolem vzniku

3 Vašák vedle úvodu vypracoval kapitoly zastřešené názvem „Teorie textologie“.

almanachu *Máj 1858*.[4] Králík se ovšem nespokojuje s výkladem pouhého svědectví, nýbrž argumentuje dalšími spojitostmi Barákem signovaného textu s povídkovou tvorbou Nerudovou, jakými jsou například ojedinělý výskyt žánru povídky v Barákově díle, tematická příbuznost, podobná lokace textu (Malá Strana a okolí).

V návaznosti na zpochybnění Barákova autorství „Kříže pod Petřínem" Králík již v následujícím roce ve článku v *Literárních novinách* upozorňuje na možnost Nerudovy literární mystifikace, když vyslovuje hypotézu o možném Nerudově autorství všech básní signovaných Barákem z přelomu padesátých a šedesátých let. V té době ovšem ještě Králík přiznává: „Nejde zatím ovšem o nic víc než o hypotézu — byť byla sebepravděpodobnější" (KRÁLÍK 1957: 6).

Tato Králíkova odvážná teze, za jejíž inspirační zdroj nutno označit zmíněnou studii M. Scherrerové, dovedla Králíka k dalšímu zkoumání materiálů, které Králík velmi hbitě zúročil v edičním výkonu, když se rozhodl vydat soubor všech Barákem podepsaných a v letech 1858–1862 ve sbornících a časopisech tištěných básní a s přívažkem „Kříže pod Petřínem". Tak vznikla ojedinělá edice *Z doby Májů*, na jejíž obálce či titulním listě sice není Nerudovo jméno uvedeno (Barákovo ovšem také ne), avšak editor svazku o jeho autorství přetištěných textů mluví již se značnou jistotou: „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).

Králíkovy atribuční pokusy vyvolaly silnou kritiku. Upozorněme zvláště na reakci Felixe Vodičky,[5] který jednak poukázal na několika příkladech na rozdílnost co do „osobnosti" Barákem podepsaných textů a Nerudovou poezií *Knih veršů*, jednak se velmi skepticky vyjádřil k možnosti mystifikace zosnované Nerudou, jež by byla dovedena k takové dokonalosti, že by Neruda v listárně redakce *Obrazů života* komunikoval s šifrou J. B. sám se sebou, že by bylo ohlašováno vydání sbírky Barákových básní tiskem, o němž by najmě sám Barák psal v letech 1862–1863 v dopisech svým přátelům! Králík se však v dalších letech dále věnoval detailnějšímu rozboru Barákových veršů a analyzoval je na pozadí Nerudovy poezie. K rozsáhlé motivicko-tematické analýze, k novému usouvztažnění všech svých dosavadních poznatků Králík dospěl ve

4 Holinové vzpomínka byla do té doby interpretována tak, že Neruda (popř. s Hálkem) měl napsat povídku „Dvojí probuzení", jež je v *Máji 1858* otištěna pod jménem K. Světlé.

5 Vodička ironicky poznamenal, že jediné, na čem se s Králíkem shoduje v jeho pohledu na Baráka, je to, „že by se na tyto verše nemělo zapomínat" (VODIČKA 1958: 6).

třetí a čtvrté kapitole své monografické práce *Křižovatky Nerudovy poezie* z roku 1965. Motivace svých výzkumů Králík odkrývá v předmluvě k této práci, v níž se ukazuje, že jeho pohled na otázku Neruda–Barák je cele podmíněn Králíkovou snahou vyložit dílo velkého českého básníka v perspektivě celé jeho vývojové logiky, která — a to platí i pro mnohé další Králíkovy práce — se jen nerada smiřuje s možnými cézurami jak na rovině geneze jednotlivých textů, tak v modelu básnické osobnosti. I proto Králík zdůrazňuje, že „básně tohoto Baráka vyplňují prázdné místo ve vývoji redaktora *Obrazů života*. A nejen to, básně, u nichž figuruje jméno onoho Nerudova přítele, jsou vklíněny do celé další tvorby Nerudovy po r. 1862" (KRÁLÍK 1965: 10).

K dalšímu oživení nerudovsko-barákovských sporů dochází s nástupem nové generace literárních vědců, konkrétně s odborným směřováním Pavla Vašáka, který na konci šedesátých let a zejména v letech sedmdesátých rozpracovává teorii určování autorství prostřednictvím nástrojů matematické lingvistiky (detailněji se na ni podíváme v následující kapitole). Králík proto znovu formuluje svůj pohled na tento problém (IDEM 1973) — nerozšiřuje ovšem nijak svoji argumentaci a otázku autorství posouvá z roviny hypotézy víceméně do roviny jistoty. Toto Králíkovo vystoupení[6] ovšem aktivizovalo k reakci Emanuela Macka, který nejprve v *České literatuře* (MACEK 1974), později v rozsáhlé studii „Králík kontra Barák", která byla otištěna ve sborníku *Literární archiv*,[7] vyvracel jednotlivé Králíkovy vývody a podrobil Králíkovy badatelské metody velmi přísné kritice. Macek tak dosud nejdůkladněji a nejsystematičtěji ověřil všechna ve sporu snesená fakta, celou otázku zasadil do širokého historického kontextu, a tak se značnou přesvědčivostí v návaznosti na dřívější Vodičkův postoj před Králíkem „obhájil" Barákovo autorství.[8]

2. Atribuce „Kříže pod Petřínem" provedená Pavlem Vašákem

Vašákova atribuce povídky „Kříž pod Petřínem" shrnutá v knize *Metody určování autorství* (1980), 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

6 K nerudovsko-barákovské otázce se Králík ovšem vyjádřil souhrnně například ještě v nedokončeném eseji *Osvobozená slova*, který byl uveřejněn ve stejnojmenném výboru Králíkových prací z roku 1995 (KRÁLÍK 1995). I v něm Králík vyslovuje svoji nedůvěru v Barákovo autorství pěti desítek básní, které — jeho optikou — geniálně anticipují pozdější Nerudův básnický vývoj (*Prosté motivy*, *Zpěvy páteční*).

7 Macek svoji studii z *České literatury* zakomponoval do příspěvku v *Literárním archivu*.

8 Králíkova reakce na Vašáka a Macka, jež byla také jeho posledním slovem k otázce Barák–Neruda, již nebyla bezprostředně zveřejněna. Z pozůstalosti ji vydal až Jiří Opelík (srov. KRÁLÍK 1993).

(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 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 nikoli 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 1 (dle VAŠÁK 1980: 190) uvádí pro text „Kříže pod Petřínem" (K) a 15 Nerudových povídek (N1–N17)[9] 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).

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 tabulky 1 (obr. 1).

9 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.

TAB. 1: 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.

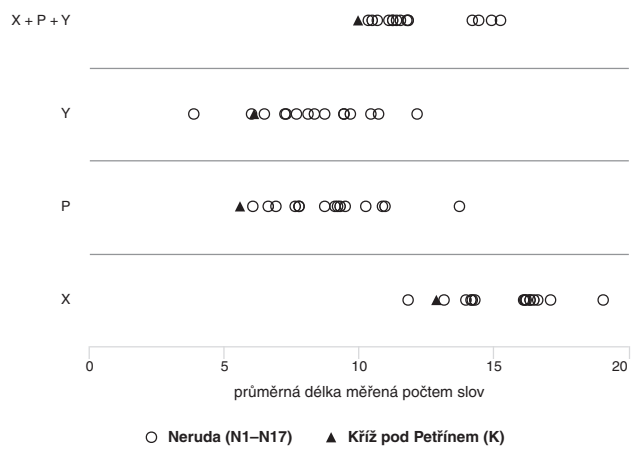
	X	P	Y	X + P + Y
N1	11,83	7,77	7,27	11,26
N2	13,96	9,33	8,74	10,52
N3	13,17	10,98	7,67	11,57
N4	16,5	7,64	9,67	14,43
N5	16,37	7,8	12,15	10,34
N6	17,13	6,05	10,75	15,26
N7	16,36	6,65	6,00	11,39
N8	16,17	10,9	8,34	11,82
N9	19,06	13,75	9,46	14,21
N11	16,66	9,14	8,12	10,69
N13	16,22	8,76	10,46	11,81
N14	14,33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

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 „Kříži pod Petřínem":

(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

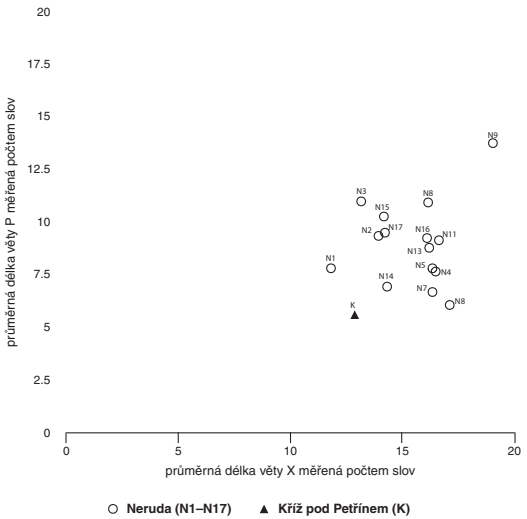


OBR. 1: Průměrná délka věty v řeči vypravěče (X), věty v řeči postav (P), uvozovací věty (Y) a průměrná délka tří předešlých typů dohromady (X + P + Y) v textu „Kříže pod Petřínem" a 15 Nerudových povídkách. Měřeno počtem slov.

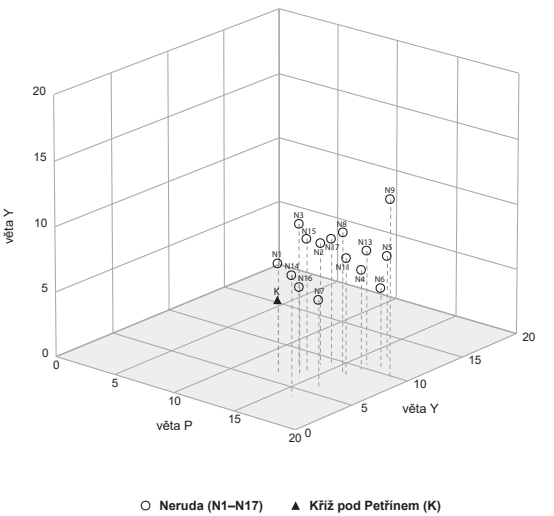
> z pěti textů nezapadajících do tohoto intervalu se vyznačují malým relativním zastoupením věty postav (tj. rozložení P), jmenovitě N4 — 19,69 %, N6 — 9,30 %, N12 — 0,00 %, N16 — 9,50 %" (VAŠÁK 1980: 77). K tomu dodejme, že s relativně nízkým zastoupením věty typu P (které má vysvětlovat anomálie u Nerudy) se setkáme i v textu „Kříže" (= 19,97; srov. IBID.: 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. 2) a následně na zobrazení hodnot zjištěných u prvních tří typů vět (X, P, Y) v trojrozměrném grafu (obr. 3). 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").

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.).



OBR. 2: 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. 3: 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.

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).[10]

(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 a s využitím robustních vícerozměrných metod současné stylometrie.

3. Kvadratická a kosinová Delta

Téměř všechny současné metody atribuce autorství jsou založeny na modelu, v němž jsou texty reprezentovány vektory určenými nějakou frekvenční charakteristikou (četnost nejfrekventovanějších slov, četnost nejfrekventovanějších bigramů aj.). Podobnost mezi jednotlivými texty je pak modelována na základě vzdálenosti mezi těmito vektory. Patrně nejrozšířenější (přinejmenším v oblasti atribuce uměleckých textů) jsou modely vycházející z tzv. míry Delta navržené Johnem Burrowsem (BURROWS 2002; 2003). Z velkého množství se zde omezíme pouze na dvě metriky vycházející z Burrowsových prací: kvadratickou Deltu (ARGAMON 2008), která opravuje některé matematické nesrovnalosti původní Burrowsovy Delty, a kosinovou Deltu (SMITH — ALDRIDGE 2011), která při srovnávacích testech vykázala nejlepší výsledky (JANNIDIS ET AL. 2015).[11]

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

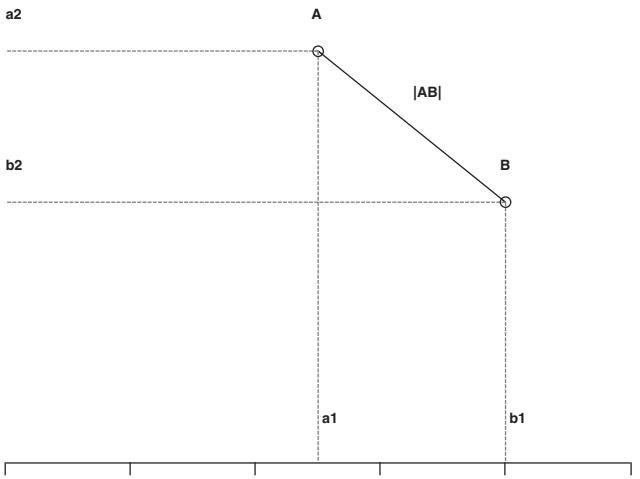
11 Ke vztahu mezi oběma metrikami viz EVERT ET AL. 2015.

3.1 Metoda

3.1.1 *Kvadratická Delta*

Vyjděme z výše uvedeného dvourozměrného zobrazení Vašákových dat (obr. 2). Krom toho, že lze vzájemnou vzdálenost mezi jednotlivými datovými body posuzovat opticky, můžeme ji i jednoduše změřit. Pro vektory (datové body) A a B v dvourozměrném prostoru, u nichž známe jak souřadnice v první dimenzi (a_1, b_1 — v tomto případě průměrná délka věty X), tak v druhé dimenzi (a_2, b_2 — v tomto případě průměrná délka věty P) se jedná o jednoduché uplatnění Pythagorovy věty — délka přepony |AB| je spočtena jako odmocnina ze součtu druhé mocniny odvěsny o délce $|a_1 - b_1|$ a druhé mocniny odvěsny o délce $|a_2 - b_2|$ (srov. obr. 4):

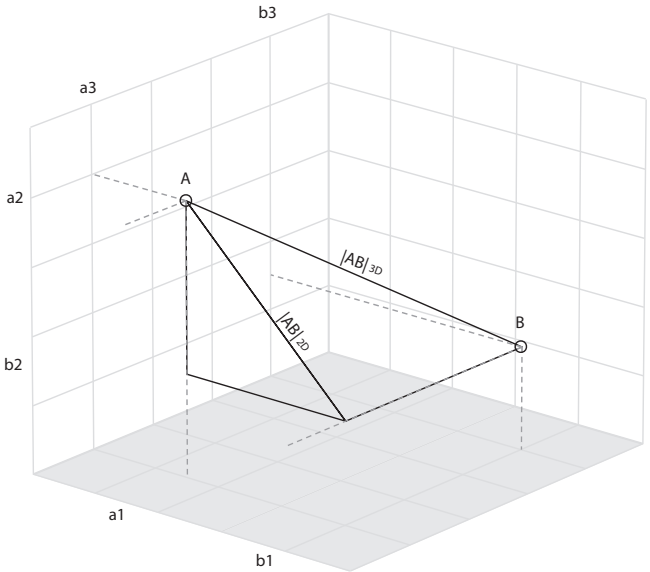
$$|\mathrm{AB}|_{2\mathrm{D}} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$



OBR. 4: Příklad výpočtu euklidovské vzdálenosti ve dvourozměrném prostoru.

Pro výpočet vzdálenosti u trojrozměrných dat (obr. 3) můžeme uplatnit týž postup. Nejdřív vypočteme výše popsaným způsobem vzdálenost v prvních dvou dimenzích ($|\mathrm{AB}|_{2\mathrm{D}}$) a následně analogickým způsobem dopočítáme vzdálenost |AB| (srov. obr. 5), tedy:

$$|\mathrm{AB}|_{3\mathrm{D}} = \sqrt{|\mathrm{AB}|_{2\mathrm{D}}^2 + (a_3 - b_3)^2} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$



OBR. 5: Příklad výpočtu euklidovské vzdálenosti v trojrozměrném prostoru.

Čtyřrozměrná data už sice přesahují možnosti lidské představivosti, snadno bychom ale došli k tomu, že vzdálenost mezi vektory A a B odpovídá:

$$|\mathrm{AB}|_{4\mathrm{D}} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + (a_4 - b_4)^2}$$

Z toho lze odvodit, že pro vzdálenost vektorů A a B v obecně *n*-rozměrném prostoru platí:

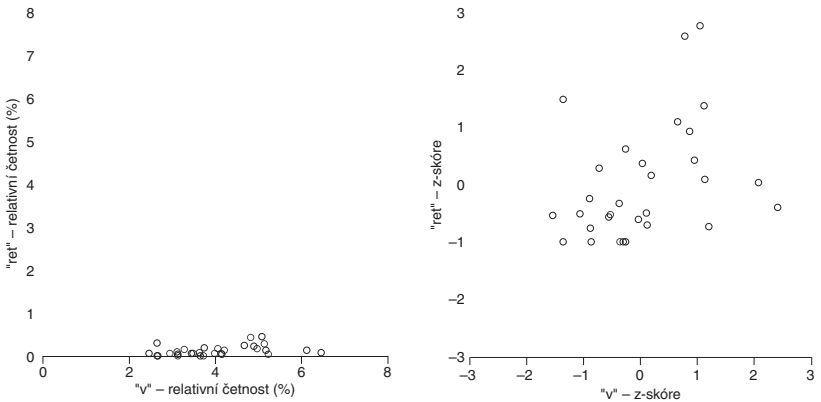
$$|\mathrm{AB}|_{n\mathrm{D}} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

Při analýze mnohorozměrných dat je pak třeba počítat s tím, že se intervaly, v nichž se zjištěné hodnoty pohybují, mohou napříč dimenzemi výrazně lišit. Markantní je to právě u vektorů určených četnostmi lexikálních jednotek. Relativní četnosti nejfrekventovanějších lemmat (typicky synsémantika jako *a*, *v*, *se*…) se budou napříč texty obvykle lišit v řádu jednotek procentních bodů. S postupem k méně a méně frekventovaným lemmatům se ale budou rozdíly mezi jednotlivými texty rapidně zmenšovat. Chceme-li bez ohledu na jejich pořadí v rankové distribuci přikládat jednotlivým lemmatům stejnou

váhu, je třeba jejich frekvence nějakým způsobem normalizovat — zjednodušeně řečeno: roztáhnout nebo smrštit intervaly, v nichž se pohybují, na stejnou délku.

Pro tyto potřeby lze využít transformaci na *z*-skóre, která upravuje původní datový soubor $\{x_1, x_2, x_3, \ldots, x_n\}$ — v tomto případě relativní frekvenci určitého lemmatu v různých textech — s průměrem μ a směrodatnou odchylkou σ na soubor s průměrem 0 a směrodatnou odchylkou 1 (srov. obr. 4):

$$z_i = \frac{x_i - \mu}{\sigma}$$



OBR. 6: Příklad: první (*v*) a sté (*ret*) nejfrekventovanější lemma v korpusu třiceti náhodně vybraných básnických sbírek. Vlevo: relativní četnosti v procentech. Vpravo: transformace těchto dat na *z*-skóre.

Právě na tomto principu je založena Argamonova kvadratická Delta (Δ_Q). Chceme-li porovnat atribuovaný text t_0 s korpusem textů $T = \{t_1, t_2, t_3, \ldots, t_m\}$ od kandidátů na jeho autorství, pak:

(1) Z korpusu T vybereme (například) n nejfrekventovanějších lemmat $l_1, l_2, l_3, \ldots l_n$.

(2) Každý text (včetně atribuovaného) reprezentujeme vektorem v n-rozměrném prostoru určeném *z*-skóry relativních frekvencí těchto lemmat v daném textu $(z_j(t_i))$:

$$\vec{t_i} = \{z_1(t_i), z_2(t_i), z_3(t_i), \ldots, z_n(t_i)\}$$

(3) Stylistickou vzdálenost (Δ_Q) mezi texty t_a a t_b spočteme jako druhou mocninu vzdálenosti mezi je reprezentujícími vektory:

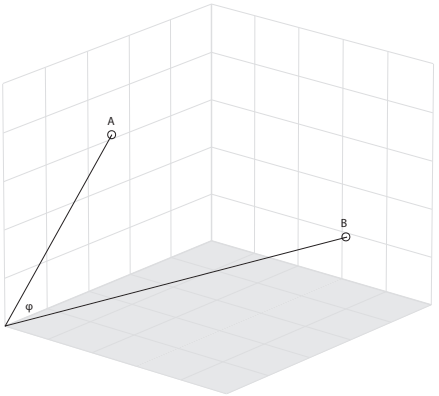
$$\Delta_Q(t_a, t_b) = |\overrightarrow{t_a t_b}|^2 = \sum_{i=1}^{n} \left(z_i(t_a) - z_i(t_b) \right)$$

Za nejpravděpodobnějšího autora t_0 je pak označen ten z kandidátů, u jehož textu byla zjištěna nejmenší vzdálenost od atribuovaného (nejbližší soused).

Je-li v korpusu T od každého autora doložen více než jeden text, můžeme zároveň jednoduše odhadnout i míru úspěšnosti zvolené metody: pro každý text $t_a \in T$ najdeme nejbližšího souseda $t_b \in T \wedge a \neq b$. Úspěšnost pak stanovujeme jako procentuální zastoupení takových párů, kde jsou t_a i t_b napsány týmž autorem.

3.1.2 *Kosinová Delta*

Smith-Aldridgova kosinová Delta ($\Delta_\angle$) funguje na stejném principu pouze s tím, že nevychází ze vzdálenosti mezi vektory, ale z velikosti úhlu, který svírají (obr. 7).[12]



OBR. 7: Příklad: úhel svíraný vektory A a B v trojrozměrném prostoru.

Kosinus úhlu svíraného vektory t_a a t_b v n-rozměrném prostoru, lze spočíst jako:

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

12 Rozdíl mezi euklidovskou vzdáleností a velikostí úhlu mezi vektory si lze nejlépe představit na příkladu hvězdné oblohy. První odpovídá skutečné vzdálenosti mezi tělesy, druhá vzdálenosti, jak se jeví pozorovateli ze Země.

Protože se jeho hodnoty mohou pohybovat mezi 1 (stejný směr vektorů) a –1 (opačný směr vektorů), je vhodnější výpočet upravit tak, aby (stejně jako u Δ_Q) odpovídala nižší hodnota větší podobnosti a naopak, proto:

$$\Delta_{\angle} = 1 - \cos\varphi$$

3.2 Výsledky

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 s celkem 38 texty autorů Barákovi časově blízkých (dále „srovnávací korpus") — krom sbírek Jana Nerudy do něj byly zahrnuty sbírky Josefa Václava Friče, Vítězslava Hálka, Adolfa Heyduka, Jiljí Vratislava Jahna, Vojtěcha Lešetického, Jaroslava Martince, Gustava Pflegera-Moravského a Václava Poka Poděbradského (soupis sbírek v tab. 3).[13] Je zřejmé, že taková analýza nemůže přímo sloužit jako argument pro Barákovo neautorství/Nerudovo autorství sporných básní, umožní nám ale učinit závěry o stylové podobnosti sporných básní a básní Jana Nerudy ve srovnání s dobovým kontextem.

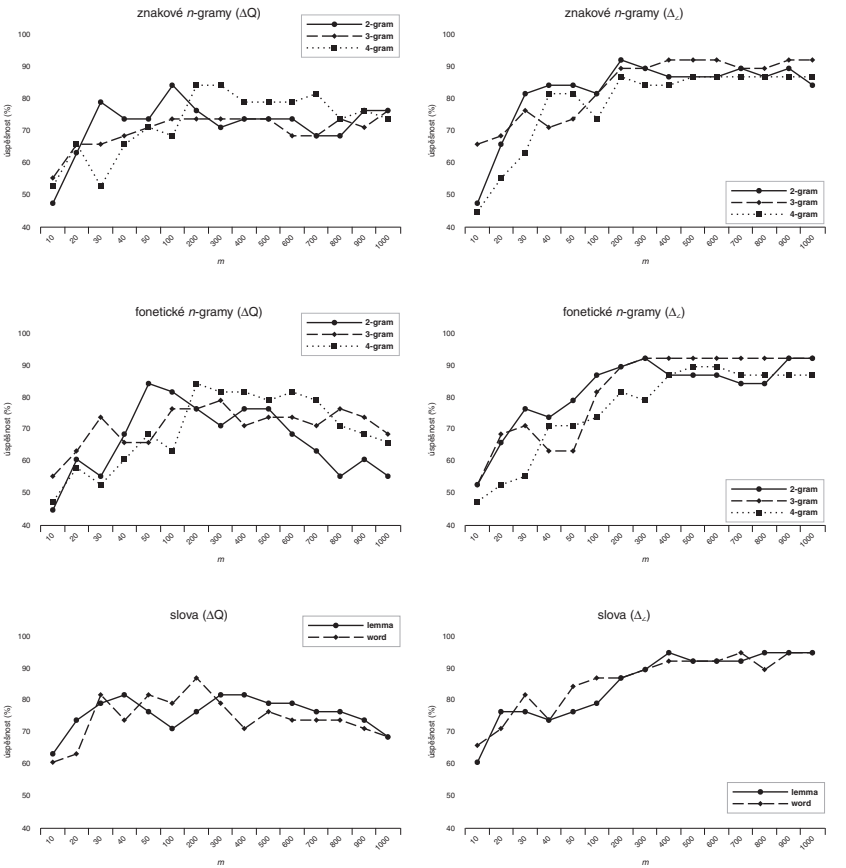
Analýzu jsme provedli na základě osvědčených stylometrických ukazatelů (četnost slovních tvarů, lemmat a znakových *n*-gramů). Protože předpokládáme, že v poezii podléhá výrazné stylizaci i zvuková stránka, připojili jsme k této sadě ještě četnosti fonetických *n*-gramů. Konkrétně jsme analyzovali:
(1) relativní četnosti *m* nejfrekventovanějších slovních tvarů;
(2) relativní četnosti *m* nejfrekventovanějších lemmat;
(3) relativní četnosti *m* nejfrekventovanějších znakových bigramů;[14]
(4) relativní četnosti *m* nejfrekventovanějších znakových trigramů;
(5) relativní četnosti *m* nejfrekventovanějších znakových tetragramů;
(6) relativní četnosti *m* nejfrekventovanějších fonetických bigramů;[15]

13 Všechny analyzované texty byly převzaty z *Korpusu českého verše* (PLECHÁČ — KOLÁR 2015). Podrobnější bibliografické informace k nim viz ‹http://versologie.cz›. Do výběru byli zahrnuti všichni autoři narození mezi lety 1825–1840, od nichž jsou v *Korpusu českého verše* v 19. století doloženy alespoň dvě různé sbírky. Z rozsáhlého díla A. Heyduka byly zařazeny pouze sbírky z šedesátých a sedmdesátých let.

14 Znakové *n*-gramy jsme zpracovávali bez rozlišení na majuskule a minuskule, včetně mezer, bez interpunkčních znaků, a to vždy v rámci jednoho verše. Např. z verše „A vidíš, kdybys měla cit" byly extrahovány následující bigramy: a | v|vi|id|dí|íš|š | k|kd|dy|yb|by|ys|s | m|mě|ěl|la|a | c|ci|it.

15 Fonetické *n*-gramy jsme zpracovávali včetně mezer zastupujících hranice mezi slovy. Např. z verše „A vidíš, kdybys měla cit" [a vďíš gdibis mňela cit] byly extrahovány následující bigramy: a | v|vi|iď|ďí|íš|š | g|gd|di|ib|bi|is|s | m|mň|ňe|el|la|a | c|ci|it.

(7) relativní četnosti *m* nejfrekventovanějších fonetických trigramů;
(8) relativní četnosti *m* nejfrekventovanějších fonetických tetragramů,
a to pro 15 různých nastavení: *m* = 10, 20, 30, 40, 50, 100, 200, 300, …, 1 000. Ve všech případech jsme měřili podobnost mezi texty jak prostřednictvím Δ_Q, tak $\Delta_<$. Celkem jsme tedy vytvořili 8 × 15 × 2 = 240 modelů. Jejich úspěšnost, tj. zastoupení sbírek ve srovnávacím korpusu, jejichž nejbližším sousedem je v daném modelu sbírka téhož autora, zobrazuje obr. 8.



OBR. 8: Úspěšnost atribučních modelů.

Je patrné, že:

(1) Na úspěšnost má zdaleka největší vliv m — u všech modelů narůstá s jeho zvyšující se hodnotou, mezi $m = 100$ a $m = 300$ se pak obvykle stabilizuje (Δ_{Z}) nebo začíná klesat (Δ_{Q}).
(2) Δ_{Z} lze považovat za spolehlivější než Δ_{Q}. U všech analyzovaných jednotek se nejlepší výsledky u Δ_{Z} pohybují mezi 80–95 % ($300 \leq m \leq 1\,000$), u Δ_{Q} mezi 70–85 %. Nadále se proto budeme zabývat pouze první zmíněnou mírou.
(3) Jednotlivé analyzované jednotky lze (přinejmenším u Δ_{Z}) pokládat za rovnocenně spolehlivé ukazatele. Znakové n-gramy, fonetické n-gramy, slovní tvary i lemmata vykazují za stejných podmínek srovnatelné výsledky.

V úhrnu tak můžeme konstatovat, že všechny modely založené na Δ_{Z} vykazují u srovnávacího korpusu minimálně od $m = 300$ vysokou úspěšnost (u 80–95 % sbírek byl autor určen správně).[16]

Teď se podívejme, které ze sbírek obsažených ve srovnávacím korpusu jsou v jednotlivých modelech vyhodnoceny jako nejbližší soused sporných básní. Tab. 2 ukazuje, že v drtivé většině z nich (103 ze 120) byla za stylisticky nejpodobnější označena táž sbírka — Nerudovy *Knihy veršů* (obsahující mimo jiné právě autorovu tvorbu z doby vzniku sporných básní). Nota bene, že modely, v nichž je za nejbližšího souseda označena jiná sbírka[17] ($m \leq 100$), patří celkově k méně spolehlivým (pouze u jednoho z nich je úspěšnost vyšší než 80 %).

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 jakýmikoli jinými sbírkami ze srovnávacího 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 ve srovnávacím korpusu obsaženy**.

16 Dodejme, že u čtyř svazků poémy „Pan Vyšínský" obsažených ve srovnávacím korpusu je to úkol poměrně jednoduchý vzhledem k tomu, že se ve všech z nich vyskytuje velké množství týchž vlastních jmen.

17 Jedná se o následující sbírky: Pro $m = 10$: J. Neruda: *Prosté motivy* (znak. 2gram, znak. 3gram, fon. 3gram), J. Neruda: *Prosté motivy* (fon. 3gram), J. Neruda: *Zpěvy páteční* (slovní tvar), A. Heyduk: *Básně* 2, 2 (lemma), G. Pfleger-Moravský: *Dumky* (znak 3gram), G. Pfleger-Moravský: *Duma* (znak 3gram). Pro $m = 20$: G. Pfleger-Moravský: *Duma* (znak 4gram, fon. 4gram), J. Neruda: *Zpěvy páteční* (slovní tvar), A. Heyduk: *Básně* 2, 2 (znak. 2gram). Pro $m = 30$ a $m = 40$: G. Pfleger-Moravský: *Duma* (znak. 4gram), A. Heyduk: *Básně* 2, 2 (lemma). Pro $m = 100$: A. Heyduk: *Básně* 2, 2 (fon. 4gram).

TAB. 2: Úspěšnost jednotlivých modelů Δ_z (v procentech, zaokrouhleno na celá čísla) ve srovnávacím korpusu. Tučně jsou vyznačeny případy, kde jsou nejbližším sousedem Barákových básní *Prosté motivy* Jana Nerudy, kurzívou jsou vyznačeny případy, kde je nejbližším sousedem Barákových básní jiná Nerudova sbírka.

jednotka	*m*														
	10	20	30	40	50	100	200	300	400	500	600	700	800	900	1000
slovní tvar	*66*	*71*	**82**	**74**	**84**	**87**	**88**	**89**	**92**	**92**	**92**	**95**	**89**	**95**	**95**
lemma	61	**76**	76	74	**76**	**79**	**87**	**89**	**95**	**92**	**92**	**92**	**95**	**95**	**95**
znak. 2gram	*47*	66	**82**	**84**	**84**	**82**	**92**	**89**	**87**	**87**	**87**	**89**	**87**	**89**	**84**
znak. 3gram	*65*	**68**	**76**	**71**	**74**	**82**	**89**	**89**	**92**	**92**	**92**	**89**	**89**	**92**	**92**
znak. 4gram	45	55	63	82	**82**	**74**	**87**	**84**	**84**	**87**	**87**	**87**	**87**	**87**	**87**
fon. 2gram	*53*	**66**	**76**	**74**	**79**	**87**	**89**	**92**	**87**	**87**	**87**	**84**	**84**	**92**	**92**
fon. 3gram	*53*	**68**	**71**	**63**	**63**	**82**	**89**	**92**	**92**	**92**	**92**	**92**	**92**	**92**	**92**
fon. 4gram	47	53	**55**	**71**	**71**	71	**74**	**82**	**79**	**87**	**89**	**89**	**87**	**87**	**87**

Jinými slovy, nějaká položka ze srovnávacího korpusu bude vždy nejbližším sousedem sporných básní, ať už je skutečný autor ve srovnávacím 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. Bootstrapovaná kosinová Delta

Boostrapování Delty (EDER 2013) bylo navrženo jako jedna z možných pojistek proti chybné atribuci v případě, kdy skutečný autor není ve srovnávacím 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 (4.1), poté výsledky, jakých dosáhla na našem materiálu (4.2).

4.1 Metoda

Při bootstrapování je provedeno k měření Delta (v našem případě $\Delta_{\angle}$) vždy s náhodným nastavením m (počet nejfrekventovanějších jednotek, které jsou analyzovány). Zjištěné vzdálenosti v rámci jednotlivých měření jsou transformovány na z-skóry.[18]

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

$$L = \bar{x} - 1{,}96\sigma$$

$$U = \bar{x} + 1{,}96\sigma$$

Při porovnání sporných básní se srovnávacím korpusem je uplatněn následující postup:

(1) Ze srovnávacího korpusu je vybrán text s nejnižší hodnotou průměru z-skórů vzdáleností od atribuované sbírky (t_1) a všechny texty t_2, …, t_n, jejichž interval spolehlivosti $\langle L_i; U_i\rangle$ se překrývá s intervalem spolehlivosti prvního textu $\langle L_1; U_1\rangle$.

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

$$c_i = \frac{U_1 - L_i}{\sum_{j=1}^{n} U_1 - L_j}$$

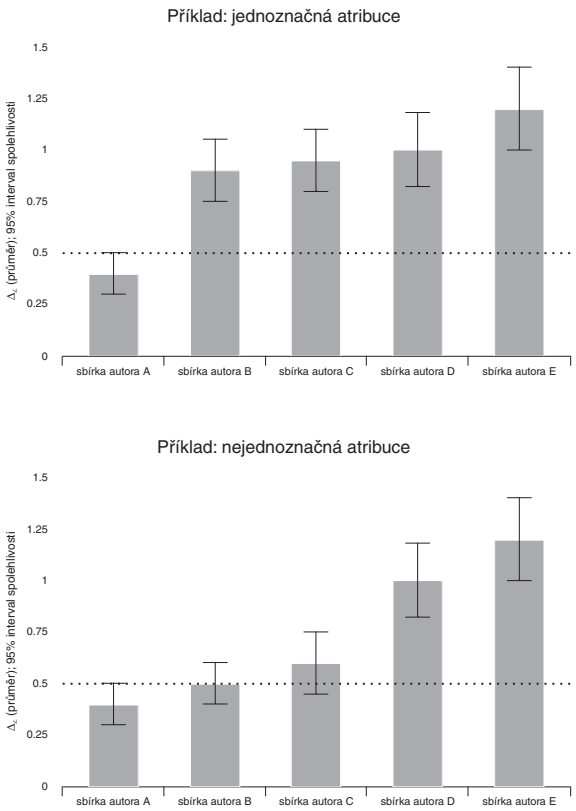
Jinými slovy, máme celkové skóre $C = 1$, které je rozděleno mezi texty t_1, t_2, …, t_n jako c_1, c_2, …, c_n tak, že

$$\sum_{i=1}^{n} c_i = C = 1$$

(3) Je-li mezi t_1, …, t_n více textů jednoho autora, je výsledné skóre pro každého autora dáno součtem skóre přidělených jeho jednotlivým textům.

18 Transformace je nezbytná u metod, kde vzdálenosti získané při různých m nejsou navzájem srovnatelné (např. Δ_Q). V případě $\Delta_{\angle}$ sice výsledky při různých m srovnatelné jsou (pohybují se vždy v intervalu $\langle 0; 2\rangle$), transformace je ale na místě vzhledem k tomu, že pracujeme s výsledky získanými na různých jednotkách (slova, lemmata, znakové a fonetické n-gramy).

Podle Ederových pozorování dochází právě v případech, kdy není skutečný autor ve srovnávacím korpusu přítomen, k tomu, že je skóre rozděleno mezi větší množství autorů.



OBR. 9: Příklad: bootstrapovaná Delta. Jednoznačná a nejednoznačná atribuce.

4.2 Výsledky

Ederovu metodu jsme aplikovali na porovnání sporných básní se srovnávacím korpusem a zároveň na porovnání každé sbírky ze srovnávacího korpusu se souborem tvořeným zbytkem korpusu a spornými básněmi. Pro každou z výše uvedených jednotek (3.2) jsme v 10 000 iteracích s náhodně nastaveným $m \in \langle 1; 1\,000 \rangle$ spočetli hodnoty Δ_z. Na základě těchto 80 000 měření jsme spočítali výsledné skóre (tabulka 3).

Pokud bychom úspěšnost metody posuzovali na základě toho, u kolika sbírek ze srovnávacího korpusu obdržel nejvyšší skóre jejich skutečný autor, dostali bychom se k bezmála 95 % (36 z 38 sbírek). Nejvíce „zmatený" byl algoritmus v případě Jahnovy sbírky *Růženec*, kde za nejpravděpodobnějšího autora označil Václava Poka, dále v případě Heydukovy sbírky *Cimbál a husle*, u níž přiřkl téměř stejné skóre jako skutečnému autorovi i Vojtěchu Lešetickému a u Lešetického *Písní a balad*, kde za nejpravděpodobnějšího autora označil naopak Heyduka (u posledních dvou jmenovaných sbírek lze důvod hledat patrně v podobné folklorní tematice, která je vzdálená jak ostatním Heydukovým sbírkám, tak druhé analyzované Lešetického sbírce *Hynek a Ráchel*).[19] Nota bene, že velice blízká skóre byla přidělena také Nerudovi a autorovi sporných básní u sbírky *Knihy veršů*. Nejpodstatnější ale je, že ve všech případech, kdy bylo jednomu autorovi přiděleno skóre 1 (15 sbírek), 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 každou sbírku s výjimkou sporných básní ještě jednou. Ze souborů, s nimiž byly sbírky srovnávány, jsme ale tentokrát odstranili texty napsané jejich skutečným autorem. Výsledky ukazuje tab. 4.

Už při letmém pohledu si lze všimnout nápadných rozdílů oproti tab. 3. Skóre jsou rozprostřena mezi mnohem větší množství autorů (v tab. 3 vykazuje u jedné sbírky nenulové skóre v průměru 1,97 autorů, v tab. 4 je to 4,87) a rozdíly mezi nimi jsou znatelně menší (v tab. 3 činí průměrný rozdíl mezi nejpravděpodobnějším a druhým nejpravděpodobnějším autorem v průměru 0,72, tab. 4 je to 0,35). Skóre 1 nebylo v tab. 4 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.

19 K tomu dodejme, že blízkost *Cimbálu a huslí* folklorní poetice — pro Heyduka atypická — byla na základě podobných metod zjištěna i na rovině výstavby verše (srov. PLECHÁČ — KOLÁR 2017: 114–117).

TAB. 3: Bootstrapovaná kosinová Delta.

	autor sporných básní	Frič	Hálek	Heyduk	Jahn	Lešetický	Martinec	Neruda	Pfleger	Pok
Frič: Různé básně		0,93							0,07	
Frič: Upír		**1,00**								
Hálek: Alfréd			**1,00**							
Hálek: Černý prapor			**1,00**							
Hálek: Dědicové Bílé hory			0,86	0,09		0,05				
Hálek: Děvče z Tater			**1,00**							
Hálek: Goar			**1,00**							
Hálek: Mejrima a Husejn			0,91						0,09	
Hálek: Večerní písně			0,76	0,19			0,05			
Heyduk: Básně	0,02			0,63				0,35		
Heyduk: Básně 2/1				**1,00**						
Heyduk: Básně 2/2				**1,00**						
Heyduk: Cimbál a husle				0,47		0,46		0,07		
Heyduk: Dědův odkaz				0,88		0,12				
Heyduk: Lesní kvítí				0,80				0,03	0,17	
Heyduk: Mohamed II				0,78		0,22				
Jahn: Růženec					0,36					0,64
Jahn: Sibylinské věštby					0,77					0,23
Lešetický: Hynek a Ráchel			0,01	0,49		0,50				
Lešetický: Písně a balady				0,77		0,23				
Martinec: Básně							**1,00**			
Martinec: Mladému pokolení							**1,00**			
Neruda: Balady a romance				0,18		0,21		0,61		

	autor sporných básní	Frič	Hálek	Heyduk	Jahn	Lešetický	Martinec	Neruda	Pfleger	Pok
Neruda: Hřbitovní kvítí	0,08			0,3			0,05	0,57		
Neruda: Knihy veršů	0,46			0,06				0,48		
Neruda: Prosté motivy	0,09							0,91		
Neruda: Písně kosmické			0,12		0,09		0,28	0,51		
Neruda: U nás	0,10							0,66	0,24	
Neruda: Zpěvy páteční				0,01	0,01			0,98		
Pfleger: Duma					0,05				0,95	
Pfleger: Dumky		0,05							0,95	
Pfleger: Královna noci				0,16					0,84	
Pfleger: Pan Vyšínský 1, 2									**1,00**	
Pfleger: Pan Vyšínský 3, 4, 5, 6									**1,00**	
Pfleger: Pan Vyšínský 7, 8, 9									**1,00**	
Pfleger: Pan Vyšínský 10, 11, 12									**1,00**	
Pok: Nová doba 1										**1,00**
Pok: Nová doba 2										**1,00**
sporné básně	—							**1,00**		

TAB. 4: Bootstrapovaná kosinová Delta, není-li mezi kandidáty zahrnut skutečný autor.

	autor sporných básní	Frič	Hálek	Heyduk	Jahn	Lešetický	Martinec	Neruda	Pfleger	Pok
Frič: Různé básně	0,06	—	0,09	0,33	0,07	0,01	0,16	0,07	0,21	
Frič: Upír		—	0,08	0,46	0,08	0,14	0,11	0,03	0,10	
Hálek: Alfréd	0,04	0,10	—	0,08	0,04	0,04	0,08	0,33	0,30	
Hálek: Černý prapor			—	0,26		0,15		0,18	0,41	
Hálek: Dědicové Bílé hory	0,01	0,07	—	0,28	0,07	0,20	0,10	0,26		0,02
Hálek: Děvče z Tater	0,04		—	0,11	0,01	0,07	0,01	0,29	0,47	
Hálek: Goar			—	0,09	0,06	0,11	0,01		0,57	0,15
Hálek: Mejrima a Husejn			—			0,03		0,13	0,84	
Hálek: Večerní písně	0,09	0,03	—	0,36			0,26	0,16	0,10	
Heyduk: Básně	0,21	0,04		—		0,11		0,64		
Heyduk: Básně 2/1	0,18	0,20	0,13	—			0,09	0,21	0,17	
Heyduk: Básně 2/2	0,18	0,25	0,19	—				0,28	0,11	
Heyduk: Cimbál a husle				—		0,87		0,13		
Heyduk: Dědův odkaz		0,09		—		0,50		0,26	0,16	
Heyduk: Lesní kvítí			0,01	—		0,16		0,20	0,63	
Heyduk: Mohamed II			0,11	—		0,89				
Jahn: Růženec					—		0,11			0,89
Jahn: Sibylinské věštby					—		0,05	0,01	0,01	0,93
Lešetický: Hynek a Ráchel		0,05	0,13	0,74		—				0,07
Lešetický: Písně a balady				0,97		—		0,03		
Martinec: Básně		0,10	0,21	0,25	0,08	0,05	—	0,27		0,04
Martinec: Mladému pokolení		0,15	0,14	0,06	0,21		—	0,21		0,23
Neruda: Balady a romance				0,47		0,53		—		

	autor sporných básní	Fričˇ	Hálek	Heyduk	Jahn	Lešetický	Martinec	Neruda	Pfleger	Pok
Neruda: Hřbitovní kvítí	0,15	0,04	0,01	0,60	0,03		0,12	—		0,04
Neruda: Knihy veršů	0,86			0,14				—		
Neruda: Prosté motivy	0,29		0,18	0,41		0,12		—		
Neruda: Písně kosmické	0,05		0,28	0,12	0,14	0,03	0,27	—	0,07	0,05
Neruda: U nás	0,16		0,02	0,07	0,08		0,09	—	0,55	0,03
Neruda: Zpěvy páteční	0,06	0,02	0,25	0,24	0,14	0,10	0,10	—	0,01	0,08
Pfleger: Duma	0,11	0,15	0,05	0,21	0,26	0,07		0,06	—	0,10
Pfleger: Dumky	0,05	0,27	0,14	0,45					—	0,09
Pfleger: Královna noci			0,17	0,80					—	0,03
Pfleger: Pan Vyšínský 1, 2			0,05	0,03	0,17			0,13	—	0,61
Pfleger: Pan Vyšínský 3, 4, 5, 6		0,03	0,80					0,15	—	0,02
Pfleger: Pan Vyšínský 7, 8, 9			0,36		0,09			0,12	—	0,43
Pfleger: Pan Vyšínský 10, 11, 12			0,56					0,28	—	0,16
Pok: Nová doba 1					0,89				0,11	—
Pok: Nová doba 2					0,60		0,19		0,21	—

5. Závěr

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) i u jednotek obsahujících informace z různých jazykových rovin (frekvence znakových a fonetických *n*-gramů). Jako pravděpodobný autor sporných básní byl pak Neruda označen i metodou citlivou k nepřítomnosti skutečného autora v souboru kandidátů (Ederova bootstrapovaná kosinová Delta).

Podobnost mohla být zapříčiněna různými důvody a v žádném případě nás neopravňuje vyslovovat zde 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. Pro posun v otázce Neruda–Barák tak na jedné straně musíme doufat jednak v možnosti dalšího upřesňujícího výzkumu jednotlivých jazykových rovin, jednak v úspěšnost v novém archivním výzkumu, který by přinesl nové mimotextové důkazy.

Stať vznikla v rámci projektu „Editologie: od náčrtu ke knize" (14–31160S), podpořeného Grantovou agenturou ČR. Při vzniku stati byly využity zdroje výzkumné infrastruktury Česká literární bibliografie <http://clb.ucl.cas.cz>.

Literatura

ARGAMON, Shlomo

2008 „Interpreting Burrows's delta: Geometric and probabilistic foundations"; *Literary and Linguistic Computing* XXIII, č. 2, s. 131–147

BURROWS, John F.

2002 „»Delta«: a measure of stylistic difference and a guide to likely authorship"; *Literary and Linguistic Computing* XVII, č. 3, s. 267–287

2003 „Questions of authorship: attribution and beyond"; *Computers and the Humanities* XXXVII, s. 5–32

EDER, Maciej

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

EVERT, Stefan — PROISL, Thomas — JANNIDIS, Fotis — PIELSTRÖM, Steffen — SCHÖCH, Christof — VITT, Thorsten

2015 „Towards a better understanding of Burrows's Delta in literary authorship attribution"; *NAACL HLT 2015 — Proceedings of the Fourth Workshop on Computational Linguistics for Literature* (Denver CO: Association for Computational Linguistics), s. 79–88

GRIEVE, Jack
2007 „Quantitative Authorship attribution: An evaluation of techniques“; *Literary and Linguistic Computing* XXII, č. 3, s. 251–270

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

JANNIDIS, Fotis — PIELSTRÖM, Steffen — SCHÖCH, Christof — VITT, Thorsten
2015 „Improving Burrows' Delta: An empirical evaluation of text distance measures“; *DH2015 Global Digital Humanities Conference* Abstracts; ‹http://dh2015.org/abstracts/›, přístup 1. 4. 2017

JUOLA, Patrick
2006 „Authorship attribution“; *Foundations and Trends in Informational Retrieval* I, č. 3, s. 233–334

KOPPEL, Moshe — SCHLER, Jonathan — ARGAMON, Shlomo
2009 „Computational methods in authorship attribution“; *Journal of the American Society for Information Science and Technology* LX, č. 1, s. 9–26

KRÁLÍK, Oldřich
1956 „Svědectví Anny Holinové“; *Host do domu* III, č. 3, březen, s. 107–109
1957 „Neruda nebo Barák?“; *Literární noviny* VI, č. 47, 23. II., s. 6
1958 „Almanach Máj 1858 a jeho redaktor“; *Slezský sborník* LVI, č. 3, září, s. 355–359
1965 *Křižovatky Nerudovy poezie* (Praha: Státní pedagogické nakladatelství)
1973 „Problém Barákova autorství“; *Česká literatura* XXI, č. 2, duben, s. 179–206
1993 „Jaro 1858“; *Sborník prací filozofické fakulty brněnské univerzity* XLII, D 40, s. 129–152
1995 „Osvobozená slova. Poznámky zestárlého literárního historika“; in idem: *Osvobozená slova*; ed. Jiří Opelík (Praha: Torst), s. 453–472 [1972/1973]

KRÁLÍK, Oldřich (ed.)
1958 *Z doby Májů* (Olomouc: Krajské nakladatelství v Olomouci)

MACEK, Emanuel
1974 „Biografická realita Josefa Baráka a jeho básnické texty“; *Česká literatura* XXII, č. 2, s. 156–170
1974a „Králík kontra Barák“; *Literární archiv PNP* VIII–IX, s. 495–580

PLECHÁČ, Petr — KOLÁR, Robetr
2015 „The Corpus of Czech Verse“; *Studia Metrica et Poetica* 2; č. 1, s. 107–118
2017 *Kapitoly z korpusové versologie* (Praha: Akropolis)

SCHERREROVÁ, Marie
1949 „Neruda a Barák“; *Slovesná věda* II, č. 2, s. 112–115

SMITH, Peter W. H. — ALDRIDGE, W.
2011 „Improving authorship attribution: Optimizing Burrows' Delta method“; *Journal of Quantitative Linguistics* XVIII, č. 1, s. 63–88

VAŠÁK, Pavel
1972 „Metody ustanovlenija spornogo avtorstva (problema Neruda — Barák)“; *Prague studies in mathematical linguistics* III (Praha), s. 143–162

1974 „Barákovo autorství jako problém“; *Česká literatura* XXII, č. 2, s. 145–155
1980 *Metody určování autorství* (Praha: Academia)
1986 *Autor, text a společnost* (Praha: Academia)

VAŠÁK, Pavel et al.
1993 *Textologie. Teorie a ediční práce* (Praha: Karolinum)

VODIČKA, Felix
1958 „Ještě jednou: Neruda nebo Barák“; *Literární noviny* VII, č. 4, 25. 1., s. 6

ŽÁČEK, Václav
1983 *Josef Barák* (Praha: Melantrich)

Summary
This article focuses on stylometric analysis of the poetic work of Josef Barák (1833–1883). The authors' introduction briefly summarizes the controversy stirred up by Oldřich Králík's theory, that Jan Neruda was the real author behind the texts, pointing out that later stylometric arguments against Neruda's authorship (by Pavel Vašák), based on an analysis of his prose work, were not actually built on solid foundations. Using quantitative analysis they then show that the poetic texts attributed to Barák show a high degree of similarity to Neruda's early collection *Knihy veršů*.

Klíčová slova/Keywords
stylometrie — atribuce autorství — textologie — Josef Barák — Jan Neruda
stylometry — authorship attribution — textual criticism — Josef Barák — Jan Neruda