Jak určit autora textu

Byli Moliére či Shakespeare skutečně autory všech jim připisovaných děl? Na podobné otázky odpovídá stylometrie, disciplína zabývající se kvantifikací stylu, nejčastěji za účelem rozpoznávání autorství.

text **PETR PLECHÁČ**

Knihovna Stylo

BURROWSOVA *Delta* i řada dalších stylometrických metod je implementovaná ve volně dostupné knihovně *Stylo* v populárním programovacím jazyku *R*. Vedle toho jsou její základní funkce dostupné i prostřednictvím uživatelsky přívětivého webového rozhraní WebSty (ws.clarin-pl.eu/websty.shtml), kde si lze základy stylometrické analýzy vyzkoušet bez jakýchkoliv vstupních předpokladů. Versologický tým Ústavu pro českou literaturu AV ČR připravuje vlastní webovou aplikaci zaměřenou na básnické texty. Ta umožní do stylometrické analýzy zahrnout nejen lexikální rysy (např. četnosti slov), ale i formální charakteristiky verše (např. četnosti rytmických konfigurací, četnosti různých typů rýmu). Aplikace bude v průběhu roku 2020 zpřístupněna na webu versologie.cz.

STYLOMETRIE patří v posledních deseti letech k nejdynamičtěji se rozvíjejícím odvětvím matematické lingvistiky. Svědčí o tom jak množství článků indexovaných ve webových databázích, množství příspěvků na odborných konferencích, tak i to, že o výsledcích stylometrického výzkumu čím dál častěji referují i běžná média (v poslední době například o studii zpochybňující Corneillovu možnou autorskou účast na hrách připisovaných Molièrovi nebo o studii potvrzující autorskou účast Johna Fletchera na hře *Jindřich VIII.*, původně publikované pod jménem Williama Shakespeara). Nejde ale – jak by se mohlo na první pohled zdát – o disciplínu etablující se teprve s nástupem moderní výpočetní techniky; její kořeny sahají hluboko do 19. století.

POČÁTKY OBORU

V roce 1851 se britský matematik Augustus de Morgan v dopise adresovaném reverendu Healdovi zamýšlel nad možnostmi určení autorství novozákonních epištol připisovaných svatému Pavlovi a navrhl, že by bylo možné odlišit texty skutečně psané Pavlem od ostatních na základě průměrné délky slov měřené počtem znaků. S povzdechem pak dodal, že *„kdyby badatelé rozuměli zákonu velkých čísel tak jako matematici, snadno by se vybralo pár set liber na to, aby se takový experiment vyzkoušel ve velkém měřítku“*.

Finance se ale podařilo sehnat až o padesát let později americkému fyzikovi Thomasi Corwinu Mendenhallovi. Ten nejprve v roce 1887 navrhl pracovat namísto průměru s celou distribucí četností slov různé délky. 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 porovnal tvar křivky určené relativními četnostmi slov různé délky v textech připisovaných Williamu Shakespearovi s křivkami extrahovanými z textů Francise Bacona a Christophera Marlowa a na základě jejich odlišnosti či podobnosti opatrně dovodil, že Shakespearovy texty nemohl napsat Bacon, zatímco je velmi pravděpodobné, že jejich skutečným autorem je Marlow. (Později se ovšem ukázalo, že celý experiment byl zkreslen jednou významnou vnější proměnnou: u Shakespeara a Marlowa zkoumal Mendenhall veršované texty, zatímco u Bacona texty neveršované.)

OD HLEDÁNÍ IDEÁLNÍHO RYSU K MULTIDIMENZIONÁLNÍM ANALÝZÁM

První polovina 20. století se nesla ve znamení hledání ideální textové charakteristiky, která v textech produkovaných jedním autorem zůstává stabilní, ale mění se napříč díly různých autorů. Vědci jich navrhli a testovali celou řadu: průměrná délka slova měřená počtem slabik, průměrná délka věty měřená počtem slov nebo různé metriky bohatosti slovníku. Žádná ale nebyla dostatečně robustní a při rozpoznávání autorství jiných textů než těch, pro něž byly původně navrženy, tyto metody obvykle selhaly.

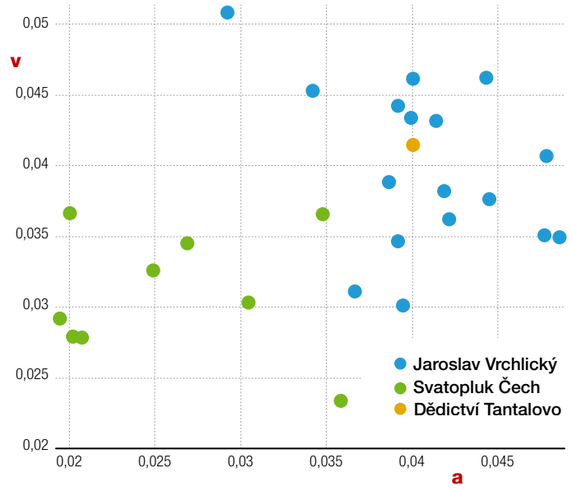
Zásadní zlom přinesla studie Fredericka Mostellera a Davida L. Wallace (1964) věnovaná autorství tzv. *Listů federalistů* (*Federalist Papers*), která patří dodnes mezi nejcitovanější práce v oboru. Autoři pracovali s nejfrekventovanějšími gramatickými slovy a rozdíly v četnostech jejich variant (např. *while/whilst*). Jejich analýza se přitom 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 (tj. srovnání jediné textové charakteristiky) k složitějším multidimenzionálním analýzám.

BURROWSOVA DELTA

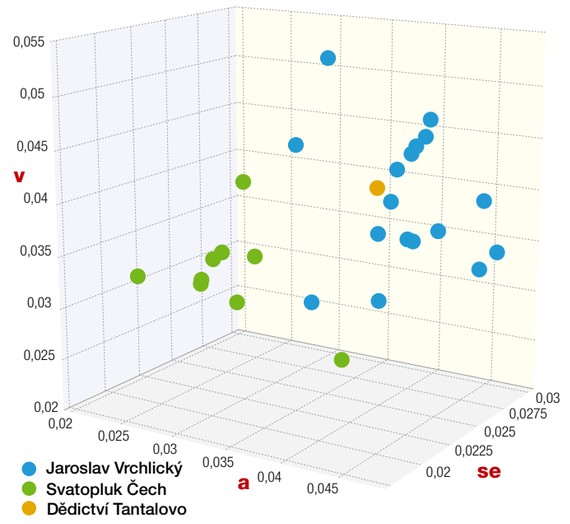
Z moderních multidimenzionálních metod se (přinejmenším v oblasti určování autorství uměleckých textů) těší dodnes velké popularitě tzv. míra *Delta*, kterou v roce 2002 navrhl John F. Burrows. Ačkoliv obvykle nedosahuje takové úspěšnosti jako složitější metody strojového učení, jde o poměrně spolehlivou a robustní metodu, postavenou na více méně intuitivních předpokladech a stojí za to se na ni podívat blíže.

Jedná se de facto o klasifikaci metodou nejbližšího souseda postavenou na četnostech nejfrekventovanějších slov. V situaci, kdy máme sporný text a uzavřenou množinu kandidátů, je tak jako nejpravděpodobnější autor označen ten, jehož texty se ve vektorovém prostoru určeném četnostmi daných slov ocitají nejblíže spornému textu. Můžeme to ilustrovat na konkrétním příkladu: předpokládejme, že básnická sbírka *Dědictví Tantalovo* se dochovala pouze v nepodepsaném rukopisu; máme ale důvod se domnívat, že sbírku napsal buď Jaroslav Vrchlický (skutečný autor), nebo Svatopluk Čech. Oba autoři vykazují určité rozdíly už v tom, jak často se v jejich sbírkách objevují pouhá dvě, respektive tři nejfrekventovanější slova – *a*, *v*, resp. *se* (**obr. 1 a 2**).

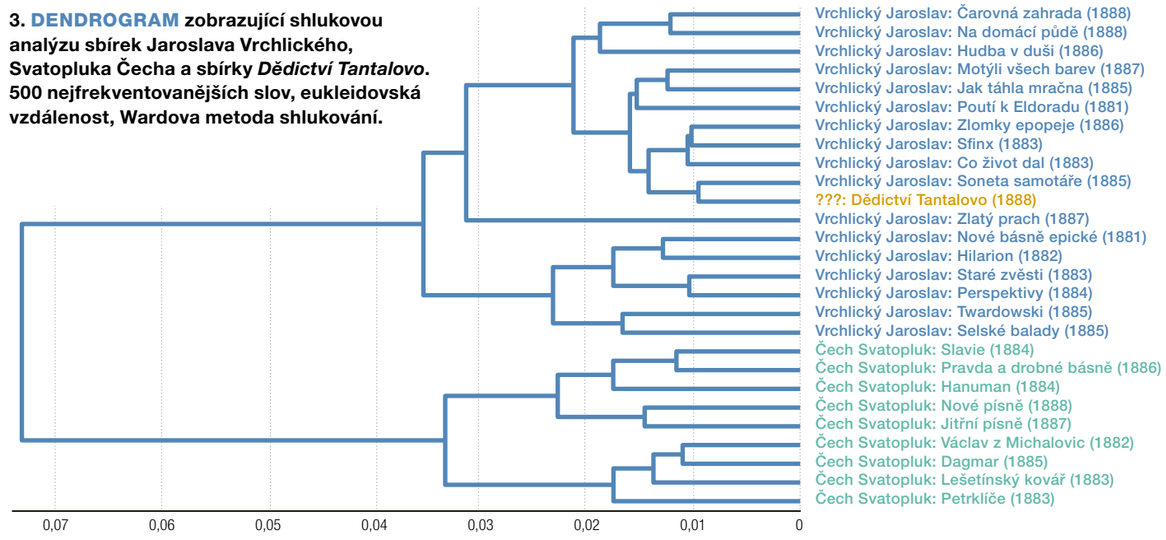
Na dvourozměrných a trojrozměrných datech lze metodu dobře ilustrovat, výsledky lze ale stěží pokládat za spolehlivé. Při řešení skutečných otázek sporného autorství se uplatňují nikoli jednotky, ale obvykle stovky až tisíce nejfrekventovanějších slov. Takové vektorové prostory už sice dalece překračují hranice lidské představivosti, vzdálenosti mezi texty (resp. jim odpovídajícími vektory) ovšem spočítat lze. Příkladem může být dendrogram, který zachycuje vzdálenosti ve vektorovém prostoru určeném 500 nejfrekventovanějšími slovy (**obr. 3**). Ten postupně spojuje texty do shluků podle jejich vzájemné (eukleidovské) vzdálenosti (osa *x*). Je patrné, že Vrchlického a Čechovy sbírky vytvářejí dva navzájem vzdálené shluky. Pokud by tedy *Dědictví Tantalovo* skutečně představovalo nepodepsaný rukopis, byli bychom schopni ho tímto způsobem spolehlivě připsat Vrchlickému.



1. RELATIVNÍ ČETNOSTI dvou nejfrekventovanějších slov (*a*, *v*) ve sbírkách Jaroslava Vrchlického, Svatopluka Čecha a sbírce *Dědictví Tantalovo*.



2. RELATIVNÍ ČETNOSTI tří nejfrekventovanějších slov (*a*, *v*, *se*) ve sbírkách Jaroslava Vrchlického, Svatopluka Čecha a sbírce *Dědictví Tantalovo*.



3. DENDROGRAM zobrazující shlukovou analýzu sbírek Jaroslava Vrchlického, Svatopluka Čecha a sbírky *Dědictví Tantalovo*. 500 nejfrekventovanějších slov, eukleidovská vzdálenost, Wardova metoda shlukování.

K dalšímu čtení…

Burrows J. F.: ‘Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing 17, 267–287, 2002/3.

Juola P.: Authorship Attribution. Foundations and Trends in Informational Retrieval 1, 233–334, 2006/3.

Koppel M., Schler J., Argamon S.: Computational Methods in Authorship Attribution. Journal of the American Society for Information Science and Technology 60, 9–26, 2009/1.

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

Mgr. PETR PLECHÁČ, Ph.D. & Ph.D., (*1985) vystudoval teorii literatury na FF UP v Olomouci a matematickou lingvistiku na FF UK v Praze. Působí v Ústavu pro českou literaturu AV ČR jako vedoucí versologického týmu. Zabývá se především kvantitativní analýzou veršovaných textů a otázkami rozpoznávání autorství.