STROJOVÉ UČENÍ V KNIHOVNĚ: VÝZVY, PROBLÉMY A PŘÍNOSY URČOVÁNÍ AUTORSTVÍ A DATACE: KORPUS ČESKÝCH AUTORŮ NA PŘELOMU 19. A 20. STOLETÍ

PhDr. Jan Pokorný, Ph.D.; jan.pokorny@nkp.cz, Národní knihovna České republiky
Mgr. et Mgr. František Válek, Ph.D.; frantisek.valek@upce.cz, Univerzita Pardubice, Fakulta filozofická
Mgr. Michal Charypar, Ph.D.; charypar@ucl.cas.cz, Ústav pro českou literaturu AV ČR
Mgr. Petr Plecháč, Ph.D. et Ph.D., DSc.; plechac@ucl.cas.cz, Ústav pro českou literaturu AV ČR
Mgr. Lenka Maixnerová; lenka.maixnerova@nkp.cz, Národní knihovna České republiky

Účel – Článek popisuje výzkum zaměřený na využití metod strojového učení pro určování autorství a datace českých textů z přelomu 19. a 20. století.

Přístup/Metody – V rámci výzkumu byly vyvinuty specializované nástroje AuthorGuesser pro identifikaci autorství a DateGuesser pro odhad časového období vzniku textu. Výzkum pracoval s korpusem digitalizovaných textů primárně české prózy od vybraných autorů z Národní digitální knihovny. Klíčové fáze technického řešení zahrnovaly přípravu a zpracování dat, delexikalizaci pro zaměření na stylistické rysy a extrakci příznaků pomocí n-gramů. Pro určování autorství byl využit klasifikátor LinearSVC. V budoucnu se plánuje významné rozšíření datasetu pro AuthorGuesser.

Výsledky – Experimenty prokázaly spolehlivost určování autorství s vysokou úspěšností i při testování na dílech kompletně vynechaných z trénování, i když s citlivostí na délku segmentů a variabilitu mezi díly. Vedlejší experiment porovnávající originál a pastiš (Jirotka vs. Macek) potvrdil schopnost modelu rozlišit autory i přes stylovou imitaci, zejména u delších textů. Naopak experiment zaměřený na dataci básnického díla Adolfa Heyduka po dobu šesti desetiletí ukázal, že určení přesného období vzniku je obtížné a jazykový styl nemusí být dostatečně stabilním ukazatelem časového zařazení na úrovni jednotek či menších desítek let. Závěry potvrzují potenciál stylometrických metod pro určování autorství, zatímco datace zůstává výzvou vyžadující další výzkum a hlubší analýzu.

Originalita/Hodnota – Výzkum se zabývá tématy dlouhodobě významnými pro obory jako literární věda a knihovnictví, jejichž důležitost roste s dostupností digitalizovaných textů. Vyvinuté specializované nástroje AuthorGuesser a DateGuesser přinášejí využití metod strojového učení pro tyto úlohy v kontextu českých textů z daného období. Výsledky potvrzují potenciál stylometrických metod v oblasti digitálních humanitních věd a knihovnictví pro určování autorství.

http://doi.org/10.52036/1335793X.2025.1.55-61

ÚVOD

Určování autorství anonymních nebo sporných textů je dlouhodobě významným tématem v literární vědě i v řadě dalších oborů, jako jsou např. knihovnictví, forenzní lingvistika, historie, archivnictví, žurnalistika, kybernetická bezpečnost, behaviorální vědy apod. Literární stylometrie, která analyzuje statistické charakteristiky psaného textu, se stala klíčovým nástrojem při identifikaci autorů, a to zejména v případech, kdy chybí přímé důkazy o původu textu. Přesné určení autorství umožňuje nejen korekci nesprávné atribuce textů, ale i hlubší analýzu stylového vývoje jednotlivých autorů a lepší pochopení literárních dějin.

S rostoucí dostupností digitalizovaných textů v knihovnách nabývá určování autorství na důležitosti i v oblasti knihovnictví. Automatizované metody umožňují s jistou měrou pravděpodobnosti analyzovat rozsáhlé korpusy, odhalovat skryté vazby mezi texty, datovat vznik děl, detekovat plagiáty či sledovat vývoj jazykových a stylistických rysů. Tyto nástroje se uplatňují při zpřesňování bibliografických záznamů, doplňování metadat nebo identifikaci neúplných či anonymních děl.

V rámci našeho výzkumu jsme vyvinuli dva specializované nástroje využívající metody strojového učení – AuthorGuesser[i] pro identifikaci autorství literárních textů a DateGuesser[ii] pro odhad časového období jejich vzniku. Pro ověření funkčnosti a spolehlivosti těchto nástrojů jsme realizovali několik typů experimentů. Hlavní část se soustředila na vývoj a testování modelu na korpusu českých textů z konce 19. a počátku 20. století, aby bylo možné prověřit, jak dobře si model poradí s autory píšícími ve stejném kulturním a jazykovém kontextu. Doplňkové experimenty měly ověřit schopnost modelu odlišit originální autorský styl od jeho napodobeniny – konkrétně jsme analyzovali rozdíly mezi textem Zdeňka Jirotky a jeho stylově imitujícími pokračováními od Miroslava Macka. Podrobnější rozbor tohoto vedlejšího experimentu je uveden v samostatné pasáži níže. Dalším cílem bylo zjistit, zda lze ze samotného jazyka a stylu rozpoznat přibližné období vzniku textu. Tento aspekt jsme zkoumali na rozsáhlém díle Adolfa Heyduka, jehož básnická tvorba se rozprostírá přes více než šest desetiletí.

URČOVÁNÍ AUTORSTVÍ LITERÁRNÍCH TEXTŮ POMOCÍ STROJOVÉHO UČENÍ

Náš výzkum se zaměřil na vývoj algoritmu pro určování autorství literárních textů pomocí metod strojového učení, konkrétně s využitím klasifikátoru LinearSVC[iii] z knihovny scikit-learn[iv]. Technické řešení zahrnovalo několik klíčových fází: výběr a příprava textů, jejich předzpracování extrakci příznaků, jejich delexikalizaci, segmentaci a následné trénování a testování modelu.

VÝBĚR A PŘÍPRAVA TEXTŮ

Projekt strojového určování autorství prozatím vychází z úzkého startovního korpusu textů, s výhledem na výrazné rozšíření v dalších fázích výzkumu. Startovní korpus byl vymezen se snahou alespoň o základní vnitřní soudržnost, pokud jde o typy zahrnutých textů, autorské osobnosti a historickou epochu, a současně o pragmatičnost s ohledem na různá omezení mj. technického rázu (kvalita OCR, nejednotný zápis metadat apod.). Vzhledem k institucionálnímu zázemí projektu korpus čerpá z textů zveřejněných v Národní digitální knihovně s volným zpřístupněním, neomezenými autorskými, dědickými nebo nakladatelskými právy. Pro korpus byly určeny a) texty v originálním českém znění, s vyjmutím překladů, kvůli důležitým lingvistickým cílům projektu; b) knižní texty, s vyjmutím periodik, kde se mísí různá autorství a kde též bývá zhoršená kvalita OCR přepisu; c) primárně texty v próze, ve většině zatím provedeného výzkumu s vyjmutím poezie, s ohledem na možnou odlišnou syntaktickou výstavbu i slovní zásobu veršované produkce.

Časový rozsah, v němž je korpus realizován, byl vymezen se zaměřením na tvorbu dvou až tří generací autorů novočeské prózy. Začátek byl stanoven v polovině 50. let 19. století, kdy už proběhly modernizační pravopisné reformy (s ohledem na vyloučení textů tištěných jednak se starším pravopisem a jednak částečně ve fraktuře) a kdy též mj. začala vycházet románová bibliotéka nakladatelky Kateřiny Jeřábkové, konec pak na polovinu 20. let 20. století, kdy se završuje tvorba řady spisovatelů a dosud se výrazněji neprojevuje činnost mladších autorů, jejichž digitalizované texty zčásti dodnes podléhají právům dědiců. Řada vybraných autorů v takto vymezeném období publikovala jak knihy beletristické prózy, tak i svazky svých publicistických článků, esejů, cestopisů, vědeckých statí apod., které jsou rovněž zahrnuty do výzkumu.

Výběr autorů probíhal podle následujících kritérií: Autor vydal v uvedeném časovém rozmezí několik knih původní české prózy, z toho vždy alespoň jednu beletristickou. Jde (s ohledem na začátek výzkumu) o autora víceméně známého, s texty, o jejichž autorství nemá literární věda pochybnosti. Do korpusu lze zahrnout jen knižní tisky vydané za autorova života, tzn. od roku 1855 po rok smrti autora včetně (cílem je vyloučit cizí textové zásahy, k nimž autor nemohl dát svolení). Z okruhu výzkumu vypadávají reedice starších textů, původně knižně publikovaných kdykoli před rokem 1855. Pro startovní korpus byla takto vyčleněna díla celkem třiceti autorů (seznam jejich knižních tisků je uveden v příslušných autorských heslech v *Lexikonu české literatury 1–4)*.

Úplný výčet použitých textů je uveden v tabulce 9 v dříve publikovaném článku Problems of Authorship Classification: Recognising the Author Style or a Book?[v]

PŘEDZPRACOVÁNÍ DATASETU

Výzkum jsme prováděli na vybraných zdigitalizovaných dokumentech získaných z Národní knihovny České republiky ve formě čistých textových souborů,

které vznikly OCR zpracováním obrazových skenů publikací. Následně jsme data upravili do jednotné podoby, konkrétně:

- Sjednocení kódování: Textové soubory byly převedeny do jednotného kódování UTF-8 pomocí balíčku chardet, který sloužil k detekci původního kódování.
- Automatické čištění: Provedli jsme ořezání čísel stránek a poznámek pod čarou, aby tyto prvky neovlivňovaly analýzu.
- Konsolidace dat: Jednotlivé textové soubory exportované z Národní digitální knihovny po stránkách byly spojeny do jednoho souboru pro každé dílo, včetně sjednocení zalomení řádků.
- Ruční úpravy: Neautorské části jako úvody, obsahy či komentáře editorů byly ručně odstraněny, aby analýza zahrnovala pouze autorský text.
- Zachování OCR chyb: Chyby vzniklé při OCR zpracování nebyly opravovány, a mohly tak ovlivnit další kroky.

EXTRAKCE PŘÍZNAKŮ

Pro extrakci příznaků jsme využili funkci CountVectorizer z knihovny scikit-learn s parametrem ngram_range=(1, 2), což nám umožnilo získat unigramy a bigramy. Například věta „Máma mele maso“ je reprezentována sadou [„máma“, „mele“, „maso“, „máma mele“, „mele maso“]. Tento přístup vychází z modelu Bag-of-Words, který reprezentuje text jako sadu slovních výrazů bez ohledu na jejich pořadí.

DELEXIKALIZACE DAT

Delexikalizace byla klíčovým krokem pro zaměření na stylistické rysy autorů místo obsahových specifik. Cílem bylo odfiltrovat prvky, které by mohly vést k rozpoznání konkrétní knihy namísto autorského stylu, jako jsou jména postav, místní názvy či archaická slova. K obohacení textů o morfologické a syntaktické informace jsme použili služby UDPipe a NameTag. Následně jsme experimentovali s různými úrovněmi delexikalizace, označenými jako „r-kódy“:

- r-04: bez delexikalizace (původní tvary slov),
- r-05: lemmatizace (použití základních tvarů slov),
- r-08: tagy slovních druhů pro autosémantická slova (např. podstatná jména, slovesa), ostatní slova byla lemmatizována, např. z věty „máma mele maso“ vzniklo „pos_noun pos_verb pos_noun“,
- další varianty (r-06 až do r-13): zahrnovaly různé kombinace tagů slovních druhů, morfologických tagů a rozpoznaných entit pomocí NameTag.

Pro finální model jsme zvolili úroveň r-08, která nejlépe vyvážila eliminaci obsahových rozdílů a zachování stylistických charakteristik.

SEGMENTACE DAT

Data jsme segmentovali pro potřeby experimentování s různými délkami textových úseků. Proces probíhal ve dvou fázích:

- Primární segmentace: Knihy byly rozděleny na segmenty po 1000 tokenech (token = slovo), přičemž jsme respektovali dělení na věty. Segmenty kratší než 500 tokenů na konci knihy byly odstraněny. Každému segmentu byl náhodně přiřazen status „train“ (60 %), „test“ (20 %) nebo „devel“ (20 %).
- Vnitřní segmentace: Segmenty o délce 1000 tokenů byly dále rozděleny na 500, 200, 100 a 50 tokenů, přičemž byly zachovány původní hranice a statusy. Segmenty kratší než polovina požadované délky byly opět vyřazeny.

Tento přístup zajistil konzistenci a srovnatelnost výsledků napříč experimenty. Vzhledem k tomu, že nebylo prováděno vylaďování modelů, základní rozdělení dat na trénovací, vývojová a testovací v poměru 60:20:20 umožnilo větší variabilitu v trénovacích a testovacích experimentech (např. využití spojených "testovacích" a "vývojových" dat pro trénování a "trénovacích" pro testování).

TRÉNOVÁNÍ A TESTOVÁNÍ MODELŮ

V první fázi jsme testovali různé algoritmy z knihovny ScikitLearn na menším vzorku šesti autorů, včetně:

- Naive Bayes (MultinomialNB),
- C-Support Vector Classification (SVC),
- Linear Support Vector Classification (LinearSVC),
- K-Nearest Neighbours (KNeighborsClassifier),
- Stochastic Gradient Descent (SGDClassifier),
- Decision Tree (DecisionTreeClassifier).

Experimenty zahrnovaly ověření vlivu délky segmentů a formy delexikalizace na výkon modelů. Pro práci s daty jsme použili knihovnu ScikitLearn (verze 1.5.1). Na základě prvotních zjištění na malém datasetu jsme dataset rozšířili na 210 knih od 23 českých autorů z konce 19. a počátku 20. století, čímž jsme zajistili větší diverzitu a robustnost modelu. Pro finální řešení AuthorGuesser jsme zvolili delexikalizaci r-08 a algoritmus LinearSVC, který vykazoval nejlepší výsledky. Modely byly natrénovány pro všechny segmentace (s-1000, s-500, s-200, s-100, s-50).

Kromě standardního testování jsme provedli i ověřující experimenty, například „leave-one-out“, kdy byl model trénován na všech knihách kromě jedné, která sloužila jako testovací. Výsledky všech testů jsme publikovali v Digital Humanities Quarterly (2023, roč. 17, č. 4), kde jsme zdůraznili nutnost oddělení trénovacích a testovacích dat na úrovni celých děl. Demonstrovali jsme, že v případě, kdy jsou k testování použity segmenty knih, jejichž jiné segmenty byly použity k trénování, pak algoritmus do velké míry nerozpoznává autora a jeho styl, ale spíše knihu samotnou.

EXPERIMENT POROVNÁVÁNÍ ORIGINÁLNÍHO AUTORSKÉHO TEXTU S JEHO PASTIŠEM

Jako vedlejší experiment jsme analyzovali texty originálního Saturnina Zdeňka Jirotky a napodobeniny Miroslava Macka, který po Jirotkově smrti napsal několik pokračování ve snaze o maximální zachování Jirotkova stylu. Cílem bylo ověřit, zda model dokáže rozlišit Jirotkovy originální texty od Mackových pastišů. Výsledky experimentu ukázaly vynikající výkonnost modelu při delších segmentech (100 % přesnost při segmentaci po 1000 tokenech), zatímco s klesající délkou segmentů přesnost postupně klesala (až na 83,23 % při segmentaci po 50 tokenech). Analýza nejdůležitějších příznaků (feature importance) odhalila, že model spoléhá na frekvenci morfologických značek (např. pos_noun, pos_verb, pos_adv) a jejich kombinací (např. pos_noun pos_verb), přičemž Jirotkovy texty vykazovaly větší váhu u slovesa „být“ (užitého jako pomocné, nikoliv plnovýznamové sloveso) a Mackovy u příznaku tag_quotes, což naznačuje rozdíly ve větné struktuře a dialogové formě. Tyto nálezy potvrzují vysokou schopnost modelu rozlišit autory i přes snahu o stylovou imitaci, zejména u delších textů. Tabulka 1 shrnuje výsledky experimentů, kdy 60 % dat od každého autora bylo použito na trénování a zbylých 40 % na evaluaci výsledků.

EXPERIMENT URČOVÁNÍ DATACE VZNIKU AUTORSKÉHO TEXTU

Druhým vedlejším experimentem, kterým jsme se zabývali, bylo určování období vzniku autorského textu. Na počátku stál předpoklad, že u autorů, kteří tvořili v delším časovém období, by mohly být pozorovány změny stylu i témat. Na pozadí těchto změn mohl být osobní vývoj autora, vliv okolního prostředí a společenských změn, inspirace jinými autory, uměleckými skupinami a literárními směry, změny v osobním životě (zejm. tragické události a rozchody mohou změnit emocionální tón děl) a někdy i snaha o inovaci a experimentování.

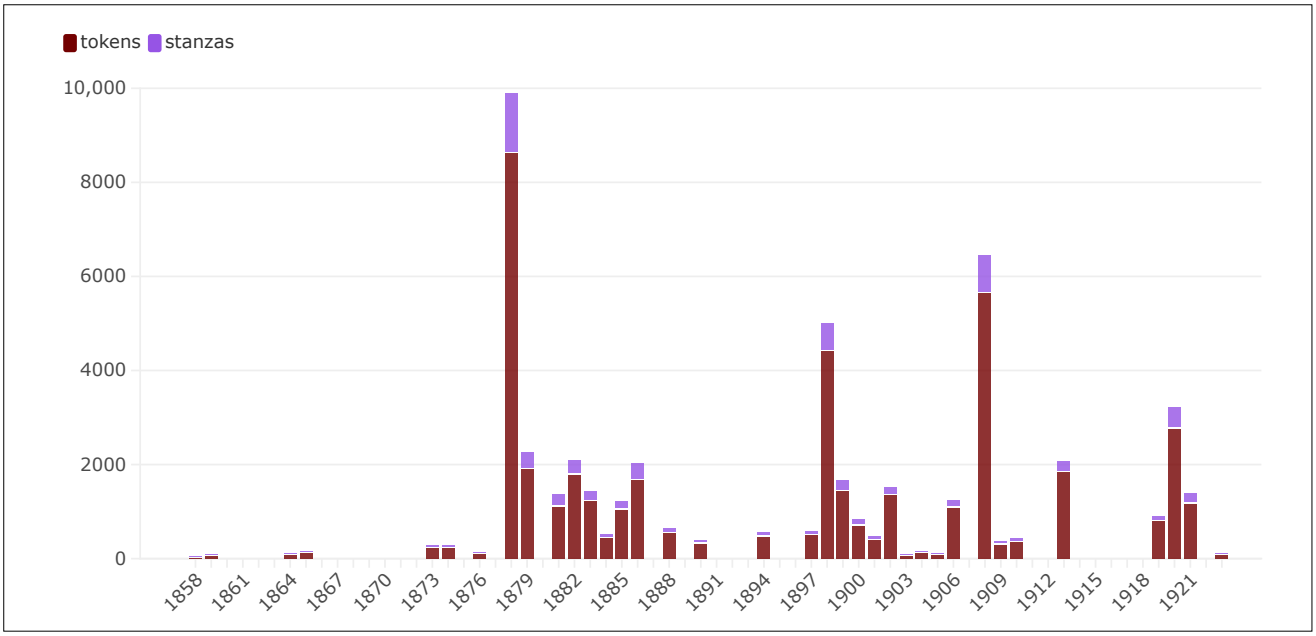
Skepticky však soudíme, že zásadní rozdíly mezi jednotlivými díly nejsou dány ani tak postupným chronologickým vývojem, jako spíše jiným tematickým zaměřením nebo žánrem konkrétního díla. Těžko také najdeme autora, ke kterému bychom měli jednotlivá časová období zastoupena rovnoměrně. K možnosti detekce data vzniku na vzorku více autorů najednou jsme pak ještě skeptičtější. Vzhledem k tomu, že předchozí experimenty ukázaly poměrně přesné detekování autorského stylu napříč déletrvajícím obdobím, soudíme, že autorský styl výrazně přehluší možná specifika užších časových období. Ostatně se nám zdá velmi nepravděpodobné, že by na úrovni jednotek let literární tvorba podléhala trendům tak, aby se tyto trendy nějak zřetelně propsaly do tvorby jednotlivých autorů, aniž by v nich později přetrvaly.

Segmentace	Počet trénovacích pasáží	Počet testovacích pasáží	Správné predikce	Nesprávné predikce	Přesnost (%)
s-1000	63	42	42	0	100 %
s-500	128	84	81	3	96,43 %
s-200	323	210	199	11	94,76 %
s-100	645	418	381	37	91,15 %
s-50	1285	829	690	139	83,23 %

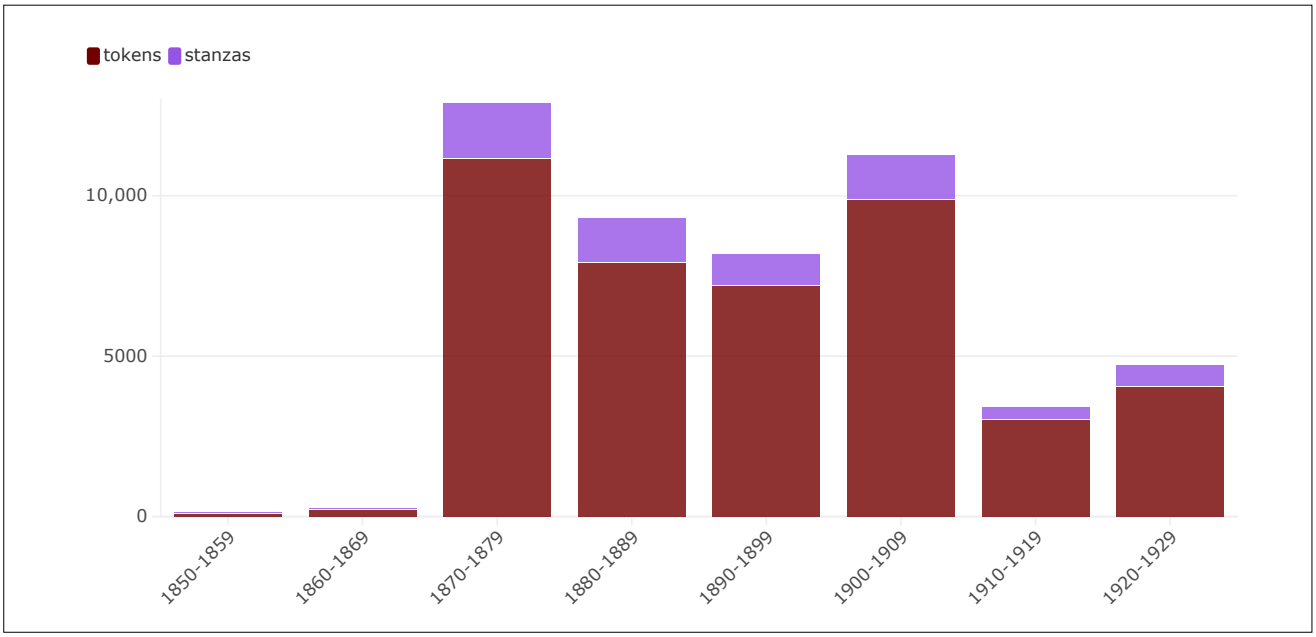
Tab. 1 Míra přesnosti podle délky segmentace

I tak jsme se rozhodli některé z našich předpokladů ověřit. Pro tento účel jsme vytvořili nový nástroj DateGuesser, který se pomocí strojového učení snažil na trénovacích datech odhadovat dobu vzniku textu. Pro experiment jsme vybrali bohaté básnické dílo Adolfa Heyduka, které čítá více než 60 básnických sbírek s převážně lyrickými básněmi s přírodní, vlasteneckou a rodinnou tematikou. Heydukovy básnické sbírky vznikaly v průběhu více než šesti desetiletí, od poloviny 19. století až do počátku 20. století. Autor byl proto vhodný z hlediska širokého časového rozpětí i tématické rozmanitosti. Heydukovy texty byly převzaty z volně přístupné mnohojazyčné kolekce básnických korpusů PoeTree[vi], budované v Ústavu pro českou literaturu AV ČR. Kompletní výčet Heydukových básní je dostupný na samostatné stránce[vii]. Texty označené jako "duplicate" nebyly do analýzy zahrnuty.

Datový vzorek jsme zkoumali rozložený po letech a po desetiletí:

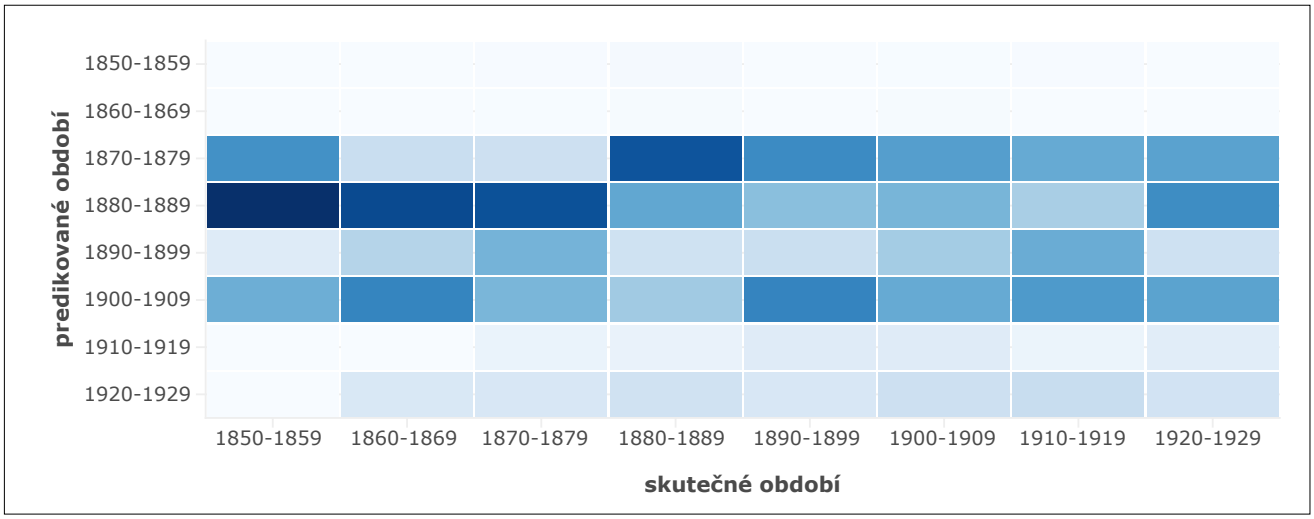


Graf 1 Rozložení vzorku po letech



Graf 2 Rozložení vzorku po desetiletích

Výsledky jsou vizualizovány v podobě heatmapy (čím tmavší, tím větší % takto tipovaných segmentů z celkového počtu segmetů pro danou periodu). Provedli jsme 73 leave-one-out-experimentů, kdy byla každá báseň jednou vynechána z trénování a použita pro evaluaci. Výsledek při segmentování po 2 strofách (protože nejmenší báseň má pouze 2):



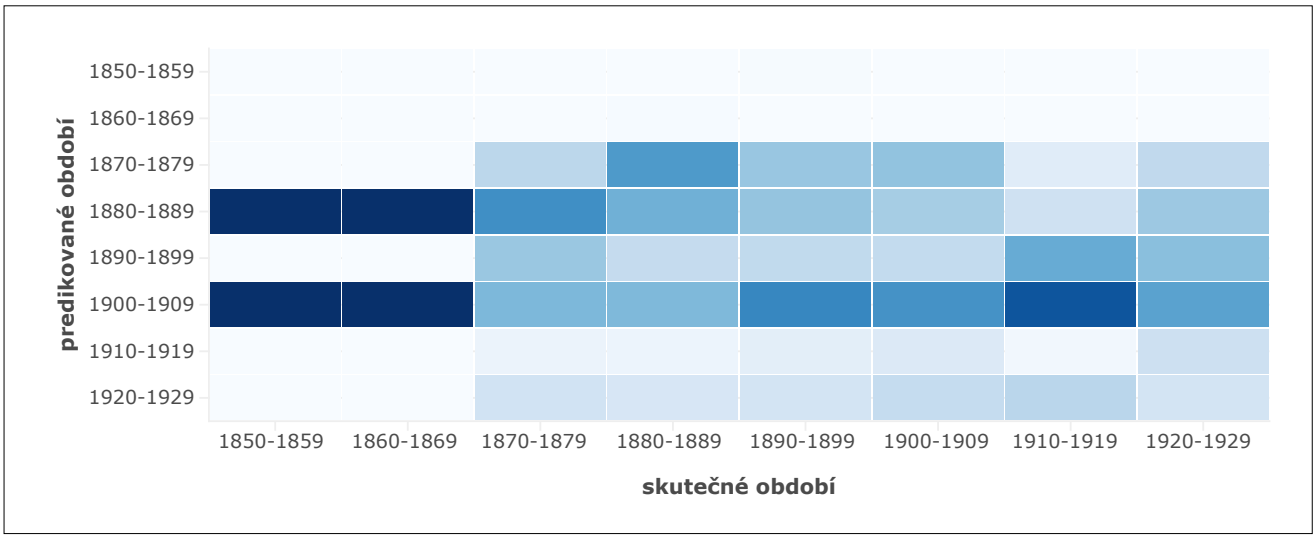
Graf 3 Míra shody časové predikce se skutečným obdobím u segmentace 2 strof

A při segmentování po 10 strofách (nebo méně v případě kratších básní) (graf 4).

Je zřejmé, že výsledky byly velmi nepřesné. A potvrdily se tak naše skeptické předpoklady, byť jen na díle jediného autora. Je pravděpodobné, že výsledky by se značně lišily dle vybraného autora. Přesto je stále možné, že u někoho by vývoj mohl být sledovatelný. Soudíme, že ani jinak koncipovaný algoritmus by v tomto ohledu neměl o moc větší úspěchy. Další experimenty v tomto směru však považujeme za neperspektivní.

VÝSLEDKY A ZÁVĚRY

Při trénování a testování na segmentech ze stejných knih dosahoval model vysoké přesnosti (např. 96 % při s-1000). Při testování na knihách kompletně vynechaných z trénování však přesnost klesla průměrně o 10–20 %, což ukazuje na význam stylistické variace



Graf 4 Míra shody časové predikce se skutečným obdobím u segmentace 10 strof

mezi díly jednoho autora. Přesto zůstává úspěšnost modelu vysoká.

Naše řešení kombinující delexikalizaci, extrakci n-gramů a LinearSVC klasifikátor prokázalo spolehlivost při určování autorství, i když je citlivé na délku segmentů a variabilitu mezi díly. Experiment s Jirotkou a Mackem potvrdil potenciál modelu v rozlišování originálu od imitace, což otevírá možnosti pro literární analýzu a digitální humanitní vědy. Naproti tomu experiment zaměřený na dataci textů ukázal, že autorský styl není vždy dostatečně stabilním ukazatelem časového zařazení – přesnost modelu byla celkově nižší, a to zejména u kratších segmentů a tematicky homogenní tvorby. Ukazuje se tedy, že zatímco určování autorství lze považovat za dobře řešitelný úkol i s relativně jednoduchými nástroji, datace textů na úrovni jednotek či menších desítek let zůstává výzvou a vyžaduje hlubší analýzu jazykových posunů i širší kontextové informace. Naše řešení potvrzuje potenciál stylometrických metod v oblasti digitálních humanitních věd, knihovnictví či forenzní lingvistiky a naznačuje směry dalšího výzkumu – například využití pokročilejších jazykových modelů, sofistikovanější delexikalizace nebo rozšíření korpusu o další žánry a jazyková období.

V následujícím období plánujeme dataset pro AuthorGuesser výrazně rozšířit (řádově o vyšší desítky autorů a se zahrnutím básnických a dramatických textů). Očekáváme, že přesnost predikce bude s narůstajícím počtem autorů klesat. To otevírá prostor pro to, abychom větší pozornost věnovali kalibraci modelů, která se v našich experimentech vyvíjí postupně.

POZNÁMKY

[i] AuthorGuesser. GitHub. Online. Praha: NK ČR. Dostupné z: https://github.com/DigilabNLCR/AuthorGuesser. [cit. 2025-06-09].

[ii] DateGuesser, (2024). GitHub. Online. Praha: NK ČR. Dostupné z: https://github.com/DigilabNLCR/DateGuesser. [cit. 2025-06-09].

[iii] LinearSVC, (c 2007-2025). Online. Scikit-learn developers. Dostupné z: https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html. [cit. 2025-06-09].

[iv] Scikit-learn, (c 2007-2025). Online. Scikit-learn developers. Dostupné z: https://scikit-learn.org/. [cit. 2025-06-09].

[v] VÁLEK, František a Jan HAJIČ, Jr., (2023). Problems of Authorship Classification: Recognising the Author Style or a Book?. DHQ: Digital Humanities Quaterly. Online. Vol. 17, no. 4. Dostupné z: https://www.digitalhumanities.org/dhq/vol/17/4/000723/000723.html. [cit. 2025-06-09].

[vi] Poetree. Online. Dostupné z: https://versologie.cz/poetree/. [cit. 2025-06-09].

[vii] Adolf Heyduk. PoeTree. Online. Dostupné z: https://versologie.cz/poetree/browser/author?corpus=cs&id_author=55. [cit. 2025-06-09].

Článek vznikl na základě institucionální podpory dlouhodobého koncepčního rozvoje Národní knihovny ČR jako výzkumné organizace poskytované Ministerstvem kultury ČR, oblast 9: Digital Humanities.

■ **Článok bol recenzovaný. / Reviewed article.**