Masarykova univerzita
Fakulta informatiky

Shrnutí charakteristických rysů autorů

Diplomová práce

Bc. Jakub Holotík

Brno, jaro 2014

Prohlášení

Prohlašuji, že tato diplomová práce je mým původním autorským dílem, které jsem vypracoval samostatně. Všechny zdroje, prameny a literaturu, které jsem při vypracování používal nebo z nich čerpal, v práci řádně cituji s uvedením úplného odkazu na příslušný zdroj.

Bc. Jakub Holotík

Vedoucí práce: RNDr. Jan Rygl

Poděkování

Předně rád bych poděkoval vedoucímu své práce, RNDr. Janu Ryglovi, za cenné rady a podněty, bez kterých by dokončení této práce nebylo možné. Zvláštní poděkování patří také mým rodičům a přátelům za podporu při studiu i tvorbě této práce a své přítelkyni za bezmeznou trpělivost.

Tato práce vznikla s podporou projektu VF20102014003 Ministerstva vnitra České republiky.

Shrnutí

Práce se podrobně zabývá historickým kontextem metod určování autorství a přináší přehled v současnosti používaných přístupů. Přínosem diplomové práce je systematický popis autorova stylomu a návrh postupu k jeho identifikaci. Součástí práce je také rozšíření systému ART o generátor shrnutí výsledků určování autorství v přirozeném jazyce.

Klíčová slova

určování autorství, stylometrie, stylom, writeprint, strojové učení

Obsah

Úvod 2
1 **Určování autorství** 3
1.1 *Definice problému* 3
1.2 *Historický rámec* 4
1.3 *Techniky kvantitativní analýzy textu* 6
1.3.1 Volba zkoumaných charakteristik textu 7
1.3.2 Výpočet hodnot měřitelných rysů dokumentu . 11
1.3.3 Určení podobnosti z hlediska autorství 12
1.3.4 Interpretace výsledků 14
2 **Charakteristické rysy autora ve stylometrii** 16
2.1 *Forenzní lingvistika* 16
2.2 *Stylometrie u soudu* 16
2.2.1 Určování autorství v českém soudnictví 17
3 **Analýza kritérií stylomu autora** 18
3.1 *Atributy dokumentů* 19
3.2 *Korpus autorských textů* 23
3.3 *Konzistence* 25
3.4 *Rozlišovací síla atributu* 26
3.5 *Jedinečnost atributu* 26
4 **Analýza řešení úkolů určování autorství** 30
4.1 *Popis systému ART* 30
4.2 *Řešení vytyčených úkolů* 33
4.2.1 Verifikace autorství 33
4.2.2 Přiřazení autora 38
4.2.3 Shlukování dle autorství 39
5 **Implementace generátoru shrnutí** 40
5.1 *Jádrový odhad hustoty (Kernel Density Estimation)* . . 40
5.2 *Použité softwarové nástroje* 43
6 **Závěr** 45
Přílohy 52
A. *Curves of style listů sv. Pavla* 52
B. *Úspěšnost charakteristik pro různé typy dokumentů* . 53

Úvod

„Kdyby byla jen jedna pravda, nebylo by možné namalovat sto obrazů na jedno téma.“
– Pablo Picaso, 1966

Určování autorství je jazykovědným problémem, se kterým se společnost potýká napříč historickými obdobími – od antiky až po současnost. S úkoly, které metody určování autorství řeší, se nesetkáváme jen v literární vědě, ale také v soudnictví, kriminalistice, žurnalistice, apod. [22].

V této práci zkoumáme z historického, lingvistického a matematikého hlediska koncept stylomu – tj. jedinečné sady charakteristických rysů autorova stylu. Tento koncept vychází z předpokladu, že jazyk každého člověka je jedinečný a je určován interakcí jedince se společností [30]. Podrobně se zabýváme také jazykovými jevy, které autorův styl mohou vystihovat.

Na základě srovnání kvalitativních a numerických metod analýzy textu, které je předmětem první kapitoly, navrhujeme teoretické předpoklady existence stylomu, kvantitativní postupy jeho identifikace a způsob integrace do systému určování autorství ART [40], který využívá technik strojového učení. Zvláštní důraz je kladen na efektivní reprezentaci a snadnou interpretaci stylomu.

Součástí této práce byla také implementace generátoru shrnutí výsledků určování autorství, kterou popisuje poslední kapitola. Práce je doplněna několika přílohami, zejména přílohou *Zpráva o výsledcích automatického určení autorství*, která je příkladem generovaného shrnutí.

1 Určování autorství

Definovat pojem autorství textu lze mnoha způsoby[1]. Český autorský zákon mluví o literárním autorském díle jako o díle slovesném vyjádřeném řečí nebo písmem, které je jedinečným výsledkem tvůrčí činnosti autora [51, § 2 odst. 1.]. Předmětem autorství je tedy nejen vnímatelná reprezentace textu[2], ale zejména jeho samotný význam. Pokud se tato práce zmiňuje o určování autorství nebo autora, vždy jde pouze o určování autorství (resp. autora) jazykového vyjádření.

Přirozený jazyk je nedílnou součástí běžné lidské komunikace. Přestože si navzájem rozumíme, jazyk každého z nás má svou jedinečnou podobu – jak z hlediska výslovnosti, tak z hlediska slovní zásoby a gramatiky. Nejvýznamnější roli při osvojování jazyka hraje jazykové a kulturní prostředí, které poskytuje zcela jedinečnou množinu učících příkladů, se kterými se v průběhu svého života setkáváme. Tato jedinečná varianta jazyka bývá označována jako idiolekt [30].

1.1 Definice problému

Rozdílný jazyk autorů je jedním ze základních východisek při určování autorství. Ač není snadné autorův specifický sloh, slovní zásobu a znalost gramatiky jednoznačně charakterizovat, právě to je cílem automatického určování autorství – totiž identifikovat takové rysy autorova stylu, které jej v dané množině autorů spolehlivě určují, a míru, s jakou k tomuto rozlišení přispívají.

Vzhledem k rozsahu pojmu určování autorství je vhodné poznamenat, že se tato práce zabývá výhradně počítačovými metodami kvantitativní analýzy textu a interpretací jejich výsledků. Následující kapitoly se soustředí zejména na tři níže uvedené problémy, které s určováním autorství souvisí:

---

1 O problematice definice autorství textu široce pojednává Harold Love ve své knize *Authorship attribution* [26, kap. 3].

2 Například strojem sázená kniha nebo ručně psaný dopis.

- **Verifikace autorství**: binární klasifikační problém

  `VSTUP:` Sada textů autora A a sada textů autora B
  `VÝSTUP:` Zda je autor A shodný s autorem B. Případným dodatečným výstupem může být také pravděpodobnost shody nebo významné rysy, které rozhodnutí vysvětlují.

- **Přiřazení autora dokumentu**: N-ární klasifikační problém

  `VSTUP:` Text neznámého autora A a korpus dokumentů N autorů – kandidátů, z nichž právě jeden je autorem A
  `VÝSTUP:` Jeden z kandidátů, popř. také významné rysy, kterými se zvolený kandidát s anonymním autorem A podobá

- **Shlukování dle autorství**: vícerozměrná shluková analýza

  `VSTUP:` Množina textů (ne nutně různých) autorů
  `VÝSTUP:` Skupiny textů, přičemž každá z nich reprezentuje právě jednoho autora.

1.2 Historický rámec

Určování autorství úzce souvisí s literárním oborem textologie, který se zabývá textem v kulturním a historickém kontextu a zkoumá proces jeho vzniku. Z historického hlediska bylo autorství textu potvrzováno (resp. vyvraceno) mnoha způsoby, např. analýzou záznamového média, grafického ztvárnění písma nebo samotného obsahu dokumentu.

Snahy o podrobné analýzy autorských děl s cílem určit autora, popřípadě také dobu a místo vzniku, lze v českých zemích pozorovat již v 18. století. V roce 1778 vydal český obrozenec Josef Dobrovský svou kritickou analýzu pražského zlomku evangelia sv. Marka [10] – tou dobou považovaného za původní rukopis Marka Evangelisty. Ve svém rozboru Markovo autorství popřel a fragment označil za pouhý opis originálu. O století později se pak rozhořela debata o pravosti rukopisů Královédvorského a Zelenohorského.

Roku 1888 Jan Gebauer ve své knize *Poučení o padělaných rukopisích Královédvorském a Zelenohorském* jejich pravost vyvrátil a za autora označil Václava Hanku, který zmíněné dokumenty v roce 1817 objevil. Rukopisy podrobil analýze technologické (tj. analýze materiálů a nástrojů), paleografické (tj. z hlediska písma), historické, zejména však analýze textologické. Zde své závěry staví především na počtu a druhu mluvnických chyb. Text o přibližné délce 6000 slov obsahuje přes 1000 mluvnických chyb, z nichž jsou mnohé pozdějšího původu, a nemohly se tak objevit v textech z 9.–14. století. [3] Obdobné jazykové chyby se zároveň nachází v Hankových textech, psaných před objevením rukopisů. Autor kritiky dále upozorňuje na myšlenky a výrazy používané až od 15. století, které však Hankovi byly známé a dostupné [13].

Gebauer ve své knize podrobně popsal charakteristické rysy jazyka sporných rukopisů, porovnal je s Hankovými dřívějšími texty a na základě podobností mu přisoudil autorství dnes již všeobecně známých falz. Stejný přístup je základem všech v současnosti používaných metod určování autorství.

Mezi nejznámější zahraniční spory (zejména v anglofonních zemích) o původ literárních děl patří spor o autorství několika divadelních her připisovaných Williamu Shakespearovi. Právě nebývalý zájem o Shakespearovo dílo podnítil řadu studií, které položily základ dnešním kvantitativním metodám. Již v roce 1787 vyslovil literární historik Edmond Malone domněnku, že ani jednu ze tří částí divadelní hry Jindřich VI nenapsal Shakespeare. Usuzoval tak proto, že na rozdíl od ostatních Shakespearových děl autor sporných textů zřídka využívá rýmu a rytmického přesahu[4]. Roku 1873 byla v Londýně dokonce založena společnost *New Shakespeare Society*, která si, mimo jiné, kladla za cíl vyřešit mnohé spory ohledně Shakespearova autorství s pomocí kvantitativních metod [16].

---

[3] „Staročesky říkalo se: *jmu se činiti něčeho*; infinitiv byl slovesa imperfektního (děje neskonalého) a jeho předmět byl v Genitivě. Spisovatelé novočeští před 1817 přejali toto rčení z jazyka starého, ale nepoznali zevrubně jeho pravidla a psali pak také: *jmu se učiniti* (perfektivně) *něco* (Akk.). ... V RKZ vyskytuje se toto několikrát, a nikdy správně, nýbrž vždycky s chybou novočeskou;“ [13, s. 26]

[4] *Enjambement* – přesahování věty nebo menšího významového celku z jednoho verše do druhého

1.3 Techniky kvantitativní analýzy textu

Metody určování autorství využívají poznatků z mnoha vědeckých oborů – od paleografie[5], literární vědy a lingvistiky až po strojové učení, statistiku, rétoriku a psycholingvistiku. Jak bylo předesláno v úvodní kapitole, tato práce se zabývá analýzou textologickou.

Cílem textologické analýzy je identifikace jedinečných vzorů jazykového chování autora. Jednotlivé techniky se pak liší způsobem, jakým jsou tyto vzory z textu získávány: zda s pomocí vlastního lingvistického instinktu a zkušenosti, nebo mechanicky specializovaným algoritmem (v dnešní době s využitím výpočetní techniky). Díky tomu můžeme metody charakterizovat podle několika kritérií. Chaski [4] uvádí následující dělení dle:

- povahy sledovaných znaků: kvalitativní – kvantitativní
- způsobu zpracování: manuálně – strojově
- typu poznatků, na kterých metoda staví: teoretické – empirické

Zatímco kvalitativní analýza se často opírá o intuici hodnotitele a vyžaduje zkušeného, pozorného čtenáře, který odhalí významné rozdíly, popřípadě důkazy o autorství, podstatou kvantitativní analýzy je předpoklad, že autora lze určit na základě porovnání několika měřitelných charakteristik textu[6].

Metody kvantitativní analýzy dnes můžeme díky výpočetní technice do značné míry automatizovat, zbavit chyb a především empiricky vyhodnotit jejich úspěšnost. Úskalím většiny dosavadních kvalitativních metod byl právě nedostatek objektivního posouzení jejich správnosti. Z historického hlediska subjektivita nutně vedla k mnoha zkreslením – každý z nás do interpretace textu nevyhnutelně vnáší své vlastní zkušenosti a očekávání. O kolizi subjektivního a algoritmizovaného určování autorství diskutuje M. Solan ve svém pojednání *Intuition versus Algorithm* [47].

---

5 *Paleografie*: Nauka zabývající se historickým vývojem písma, jeho podobami a nástroji používanými k jeho zaznamenávání

6 Nejčastěji frekvence určitého lingvistického jevu – například gramatických chyb (jako u Gebauera [13]), nebo rýmu v dříve zmiňovaném spor o autorství Shakespearovy hry Jindřich VI.

Vymezme nyní zásadní úkoly kvantitativní analýzy, kterým na následujících stránkách budeme věnovat:

- Volba zkoumaných charakteristik textu
- Výpočet hodnot měřitelných rysů dokumentu
- Určení podobnosti z hlediska autorství
- Interpretace výsledků

1.3.1 Volba zkoumaných charakteristik textu

Při kvantitativní analýze je klíčová zejména volba dostatečně určující množiny charakteristických znaků a výběr relevantních porovnávaných textů.

Uveďme stručný popis charakteristik, které používá Rygl [40, 41] v systému ART[7]; této aplikaci, která slouží k určování autorství, je věnována podkapitola 4.1:

i. **Průměrná délka vět**
Podobně jako průměrná délka slov [31], tak i průměrná délka vět (ať už počítáme písmena: Mascol [28], nebo slova: Sherman [44]) patří mezi jedny z nejčastěji užívaných charakteristik. Přestože se jako jedinečný ukazatel autorova stylu tato metrika příliš neosvědčila, v praxi je v kombinaci s dalšími charakteristikami běžně používána [16].

ii. **Frekvence celých morfologických značek**
Morfologickou analýzou se zabývá většina technik automatického zpracování přirozeného jazyka a nejinak je tomu při určování autorství českých textů. Tato metoda spočívá ve stanovení relativní četnoste celých morfologických značek všech slov analyzovaného dokumentu.

[7] *ART*: Authorship Recognition Tool [40]

iii. **Frekvence morfologických kategorií**
Morfologické značky, zmiňované v předchozí metodě, jsou složeny z různého počtu značek – kategorií, které jsou pro dané slovo (resp. slovní druh) významné.

Výstupem této metody jsou relativní četnosti různých morfologických kategorií; jejich výčet lze nalézt v dokumentaci analyzátoru Ajka [43].

iv. **Frekvence slovních druhů**
Dělení slov dle slovního druhu je jedním z nejzákladnějších lingvistických dělení a je součástí morfologické analýzy. Metrika navrhovaná Ryglem uvažuje 14 slovních druhů (10 tradičních a 4 speciální) a počítá jejich frekvenci v dokumentu.

v. **Frekvence bigramů slovních druhů**
Metoda staví na předpokladu, že zejména v jazycích s poměrně volným slovosledem (například v češtině) může být preference určitého pořadí slov (resp. různých slovních druhů) ve větě silným charakteristickým znakem autorova stylu. Tato metrika je založena na výpočtu frekvencí bigramů slovních druhů.

vi. **Frekvence stop slov**
Za stop slova v tomto kontextu označujeme lemmata n nejčastěji užívaných slov daného jazyka. Hlavním přínosem této metody je její žánrová nezávislost (za předpokladu, že seznam nejčastějších slov vznikl z žánrově vyváženého korpusu).

vii. **Opakování slov**
Opakování stejných slov bývá často ukazatelem nedostatečné slovní zásoby a obecně nízké jazykové úrovně autora – ať už v mluveném nebo psaném projevu [48]. Metoda užívaná v systému ART sčítá frekvenci, s jakou se v textu opakují slova různých slovních druhů.

viii. **Bohatost slovní zásoby**
Tato charakteristika, konkrétně číselné vyjádření bohatosti autorovy slovní zásob, byla předmětem zájmu mnohých jazykovědců. Jedním z průkopníků tohoto výzkumu byl G. U. Yule, který hodnotil individuální slovník autora pomocí své míry K,

zvané krátce *The Characteristic*, založené podobně jako předchozí metoda na opakování slov [49] (def. 1.1).

Podobných metrik bylo navrženo mnoho a jejich přehledný výčet uvádí Grieve [16, s. 24]. V dnešní době nejpoužívanější mírou tohoto typu je Simpsonovo D (*Measurement of Diversity*), které měří pravděpodobnost, že dvě náhodně vybraná slova z textu budou stejná [46] (def. 1.2).

$$K = 10^4 \left(\frac{\sum i^2 \cdot V_i - N}{N^2} \right) \tag{1.1}$$

$$D = \frac{\sum V_i \cdot i \cdot (i-1)}{N \cdot (N-1)} \tag{1.2}$$

V_i: počet slov, které se v textu vyskytují i krát
N: celkový počet slov v textu

ix. **Syntaktická analýza**
Zatímco se většina tradičních charakteristik užívaných při určování autorství zabývá lexikálními jevy, tato metoda zkoumá syntaktickou strukturu vět. Syntaktickou stylometrií a jejími metodami se zabývá např. Hollingsworth [20].

Použitelnost této metody závisí na dostupnosti a přesnosti syntaktického analyzátoru. Pro český jazyk je možné k syntaktické analýze použít například nástroj SET [24].

Navrhovaná charakteristika autora počítá frekvenci různých syntaktických značek ve struktuře textu a četnost vybraných vlastnosti syntaktického stromu jako hloubku, počet uzlů ve stromě apod.

x. **N-gramová syntaktická analýza**
Tato charakteristika pracuje s relativními počty syntaktických n-gramů (uspořádaných n-tic po sobě jdoucích navzájem závislých uzlů syntaktického stromu), je tedy rozšířením předcházející charakteristiky. Metoda staví na předpokladu, že sled funkčních částí věty je dán individuálními preferencemi autora.

xi. **Velká písmena**
Použití velkých písmen je jedním z jazykových rysů, které je možné sledovat pouze v psaném projevu. V rámci této metody je určována frekvence slov tří různých tříd:

- slova psána celá velkými písmeny
- slova psána celá malými písmeny
- slova, jejichž jediným velkým písmenem je to první

xii. **Frekvence interpunkčních znamének**
Použití některých interpunkčních znamének (případně jejich kombinací) může být při srovnávání žánrově velmi podobných dokumentů významným ukazatelem autorství textu. Metoda, kterou používá aplikace ART, sleduje frekvenci interpunkčních znamének v různých částech věty (na začátku, uprostřed a na konci).

xiii. **Typografické chyby**
Tato metrika počítá relativní četnost (vzhledem k počtu znaků) různých typografických chyb v textu (např. zdvojení nebo vynechání mezery).

xiv. **Užití emotikon**
ART hledá v textu výskyty 128 různých emotikon (28 pozitivně, 42 negativně zabarvených a 58 neutrálních) a počítá jejich frekvenci. Zároveň je uvažována také frekvence tří popsaných skupin podle zabarvení.

Většina uvedených charakteristik nebyla původně navržena pro flektivní jazyk jako čeština – výpočet některých metrik je proto pro české dokumenty výrazně složitější, než např. pro anglické. Nejzásadnějším problémem, který musí systémy určování autorství českých textů překonat, je morfologická analýza.

Lze očekávat, že statistická významnost měřených znaků je závislá na délce a typu porovnávaných textů. Zatímco dlouhé texty poskytují kvantitativním metodám dostatek informací o vzorech jazykového chování autora, u kratších textů nelze některé z naměřených charakteristik prohlásit za neměnnou součást autorova stylu.

To potvrzuje i studie [41]. Autor v ní své závěry staví na analýze metod určování autorství těchto pěti typů dokumentu (seřazených od nejdelšího po nejkratší):

- Knihy
- Články na blogu
- Příspěvky v diskuzním fóru
- Komentáře k internetovým článkům
- Tweety - příspěvky na sociální síti Twitter[8]

Stejné kategorie textů využívá také systém určování autorství popisovaný v této práci. Přesnost kvantitativních metod – stejně jako úspěšnost jednotlivých charakteristik – je obdobně závislá na počtu srovnávaných dokumentů. Čím více jich má analýza k dispozici, tím jsou její závěry přesnější [27].

1.3.2 Výpočet hodnot měřitelných rysů dokumentu

Před příchodem počítačových technologií bylo nutné hodnoty všech měřených charakteristik počítat ručně – taková analýza byla velmi zdlouhavá, náročná a náchylná k chybám.

T. C. Mendenhall v roce 1887 navrhl jednu z prvních univerzálně použitelných kvantitativních metod určování autorství, porovnávající frekvenční rozložení slov různé délky ve zkoumaných textech. Později touto metodou zpracoval téměř celé Shakespearovo dílo – manuální analýzu necelých dvou milionů slov společně zpracovávaly dvě najaté pracovnice po dobu několika měsíců, vždy nanejvýš 3 až 5 hodin denně [32]. Autor v závěru své zprávy o provedeném výzkumu dodává, že míra délky slov není pro potřeby určování autorství dostatečně spolehlivá.

Pro úplnost uveďme také další pionýry kvantitativní analýzy, zejména W. B. Smithe, který pod pseudonymem Conrad Mascol [39]

[8] Twitter: `http://www.twitter.com`

vydal v roce 1888 svou kritiku Pavlových epištol a L. A. Shermana, který ve stejném roce uvedl svou metodu určování autorství na základě průměrného počtu slov ve větě [44]. K některým z těchto metod se budeme v dalším textu vracet, protože jsou významné jak z hlediska historie, tak stylometrie.

S rostoucím počtem možných autorů, dostupných textů a měřených charakteristik se ruční analýza stává prakticky neproveditelnou a moderní kvantitativní metody se bez výpočetní techniky neobejdou.

1.3.3 Určení podobnosti z hlediska autorství

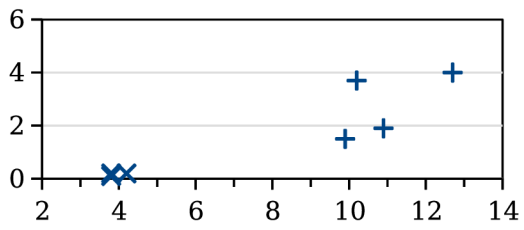
Hodnoty měřených atributů upozorňují na jazykové jevy v daném textu, o autorovi samotném však bez srovnání příliš nevypovídají. Při analýze sporných dokumentů si analytik klade otázky jako:

- *„Jak je text pro svého autora typický?“*
- *„Čím se autor textu vyznačuje a liší od ostatních?“*
- *„Mohl text napsat někdo jiný?“*

Existuje mnoho technik, které nabízí odpovědi na tyto otázky, pro jednoduchost je v následujícím textu rozdělme do tří tříd podle povahy odpovědi, kterou uživateli dávají:

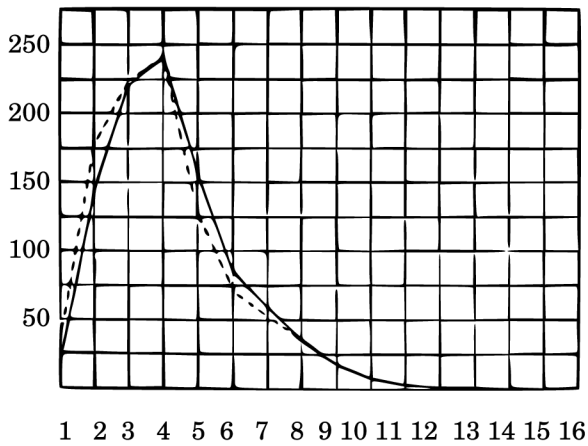
- **Vizuální**

  Triviálním příkladem grafické analytické metody jsou bodové grafy v nízkých dimenzích – shluky v bodovém grafu ukazují na texty, pro které nabývá daný atribut podobných hodnot.



Obrázek 1.1: Bodový graf šesti Pavlových epištol

Výše zmíněná Mendenhallova metoda charakteristických křivek stavěla své závěry na grafickém znázornění četností slov dané délky v textech. Na obrázku 1.2 autor srovnává křivky dvou skupin Shakespearových prací – poezie (básně *Znásilnění Lukrecie* a *Venuše a Adonis*) a ostatní prózy.



Obrázek 1.2: Charakteristické křivky pro Shakespearovu poezii a prózu (přerušovaně) [32, s. 103]

Mascol se v kritice Pavlových epištol zabývá 40 různými jazykovými ukazateli[9], jejichž hodnoty zanáší jako body do grafu. Tvar výsledné křivky, Mascolem označované jako *curve of style*, je podle něj jedinečný každému autorovi a svědčí o tom, že epištoly přisuzované sv. Pavlovi mají nejméně dva různé autory. Rekonstrukce Mascolových křivek je předmětem přílohy A. [28, 29].

- **Statistické**
  (*a*) Cílem jedné skupiny statistických metod je určit autora (popřípadě verifikovat autorství) dokumentu porovnáním s texty ostatních autorů. Mezi tyto metody patří Chi-square test nebo analýza rozptylu. (*b*) Druhá skupina pak vymezuje ze sady jazykových jevů ty, které jsou pro autora určující. Jde například

[9] Průměrná délka vět, četnost některých slovních druhů, interpunkčních znamének, morfologických tvarů, apod.

o shlukovou nebo lineární diskriminační analýzu. Vyčerpávající seznam statistických technik používaných při určování autorství sestavil Juola [22, kap. 5].

- **Techniky strojového učení**
  Z pohledu strojového učení je určování autorství specializací problému klasifikace textu. Každý dokument je reprezentován uspořádanou n-ticí binárních, celých nebo racionálních čísel – zpravidla hodnot různých atributů. Cílem metod strojového učení je pak najít klasifikační funkci, která libovolnému dokumentu (n-tici) přiřadí autora (popřípadě nepřiřadí žádného). Existuje mnoho metod strojového učení, které lze k určování autorství využít – jejich podrobný výčet uvádí Koppel et al. [23]. Pro obecný přehled jmenujme například neuronové sítě, Naivní Bayesův klasifikátor[34, kap. 1] a Support Vector Machines[21].

1.3.4 Interpretace výsledků

Jak se ukázalo například u expertních systémů, samotná odpověď na zadanou otázku je často vnímána jako irelevantní, pokud není žádným způsobem podáno její vysvětlení [36, 18] a to i tehdy, je-li úspěšnost použitého systému obecně považována za dobrou[37].

Výstupem aplikací určování autorství jsou, mimo samotné odpovědi, hodnoty měřených atributů napříč zkoumanými dokumenty. Ty jsou zpravidla uváděny jako dvojice `atribut` – `hodnota`. Vzhledem ke své povaze je tento formát vhodný k následnému počítačovému zpracování, pro člověka je však jen s obtížemi interpretovatelný. Mají-li být výsledky automatického určování autorství vyhodnocovány člověkem, jejich efektivní prezentace nabývá na důležitosti. Významným kritériem při výběru analytické metody je proto povaha očekávané odpovědi.

Vizualizace uvedená výše je jedním z běžně používaných způsobů, jakými jsou numerické výsledky předkládány k vyhodnocení. Jak však ukazuje Portet et al. [36], vizualizace je bez doprovodného (byť i automaticky generovaného) popisu často nedostatečná. Cílem této práce je proto navrhnout a implementovat takové metody, jimiž

bude možné výsledky určování autorství vyjádřit v přirozeném jazyce a usnadnit tak jejich vyhodnocování.

Podobné snahy lze pozorovat v mnoha odvětvích – lékařství [36], meteorologie [38] nebo geografie [50]; v oblasti určování autorství však autorům není znám žádný souběžný výzkum generování výstupu v přirozeném jazyce.

2 Charakteristické rysy autora ve stylometrii

Stylometrie je věda, která se zabývá analýzou individuálních odchylek v použití psaného jazyka. Cílem studia individuálního jazyka je najít a popsat jeho charakteristické rysy. Mluvíme-li o autorově stylu, máme na mysli některé (popřípadě všechny) vzory jeho jazykového chování – ať už jsou jedinečné, či běžné [8].

2.1 Forenzní lingvistika

Forenzní lingvistika je aplikovaným jazykovědným oborem, který se zabývá jazykem v kontextu práva a kriminologie. Popis vývoje a významných milníků tohoto odvětví lingvistiky spadá mimo rozsah této práce, a čtenáře proto odkazujeme například na knihu *Forensic Linguistics: Advances in Forensic Stylistics* Geralda McMenamina [30]. Forenzní lingvistika a stylometrie se zpravidla prolínají tehdy, kdy je zapotřebí určit autora sporných dokumentů, popřípadě jen jeho přibližný věk, národnost, vzdělání, pohlaví apod. – ať už za účelem kriminalistického vyšetřování, nebo dokazování v soudním procesu.

2.2 Stylometrie u soudu

Stylometrická analýza hrála významnou roli v několika zahraničních soudních sporech. Níže uvádíme jejich stručný přehled a odkazy na příslušné rozsudky a znalecké posudky.

Mezi jeden z nejexponovanějších soudních procesů, ve kterých byly využity metody určování autorství, patří spor „Ceglia v. Zuckerberg & Facebook Inc.“ o podíl ve zmíněné společnosti. Jde také o jeden z mála soudních procesů, využívajících určování autorství, jehož dokumentace je veřejně přístupná. Autorem znaleckého posudku je významný forenzní lingvista Gerald McMenamin.

Ten v něm vyvrací Zuckerbergovo autorství zkoumaných dokumentů. Svou analýzu dokládá odlišnostmi v použití následujících stylometrických znaků:

- **Interpunkce**: frekvence užití apostrofů
- **Interpunkce**: frekvence použití typografické výpustky (...)
- **Hláskování**: „back end“ vs. „front end“
- **Hláskování**: „Internet“ vs. „internet“
- **Hláskování**: „cannot“ vs. „can not“
- **Syntax**: skládání dvou vět bez použití interpunkce
- **Sytanx**: citoslovce na začátku věty
- **Syntax**: frekvence „sorry“ jako prvního slova věty
- **Syntax**: vzdálené nebo nejasné zájmenné odkazy
- **Syntax**: vynechání čárky za vedlejší větou podmiňovací
- **Užití stop slov**: frekvence užití „Thanks“ na konci dokumentu

Znalecký posudek i výsledný rozsudek, který bere expertízu ve forenzní lingvistice v úvahu, je dostupný na serveru Justia [3].

Dalším případem je proces vraždy „Commonwealth vs. Richard Wayne Illes Sr.“ [5]. Agent F.B.I. Fitzgerld ve svém posudku pro tento případ srovnává dopisy obžalovaného s jeho známými texty. Tato analýza byla zahrnuta do rozsudku jako přípustný důkaz (v odkazovaném dokumentu nejsou srovnávané jazykové rysy blíže uvedeny).

2.2.1 Určování autorství v českém soudnictví

K určování autorství na základě stylometrické analýzy nebývá u českých soudů příliš přihlíženo a autorům není znám žádný rozsudek, který by se do značné míry opíral o znalecký posudek v oblasti stylometrie.

3 Analýza kritérií stylomu autora

„Autor ke slovu se má stejně jako malíř k barvě – ty totiž představují vše, co mohou použít k naplnění svého smyslu – tvorby.“
– George Udny Yule, 1944

V úvodní kapitole jsme představili teorii idiolektu[1], pro potřeby shrnutí charakteristických rysů autora však potřebujeme jeho specifický styl systematicky popsat. Myšlenka jednoznačné, kvantifikované charakterizace jazyka jednotlivce se objevovala již v 19.století. Mezi prvními pokusy o takový popis psaného projevu uveďme následující metody a jejich autory (v chronologickém uspořádání):

- Fleay, *Shakespeare Manual*, 1876 [11, s. 69-72]

  Autor knihu vydal jako souhrn Shakespearova díla, na popud *New Shakespeare Society* v ní však uvádí také seznam několika metrických[2] pravidel, která podle něj Shakespeara jednoznačně odlišují od ostatních autorů žijících ve stejném období. Tato práce bývá často kritizována zejména pro subjektivitu, nepřesnost vstupních dat a její přílišnou konkretizaci [16].

- Mendenhall,*The Characteristic Curves of Composition*, 1887 [31]

  V době, kdy již dříve zmíněnou metodu charakteristických křivek Mendenhall publikoval, byl přesvědčen, že tvar grafu četností slov dané délky je – jak již z názvu může být patrné – charakteristický a jedinečný každému autorovi. Sám autor byl však později nucen uznat, že jeho technika není pro potřeby určování autorství spolehlivá a slouží lépe jako ukazatel jazyka dokumentu. Přesto ve výzkumu neustal a metodou průměrné délky slov se zabýval i později [32].

---

1 *Idiolekt* – varianta přirozeného jazyka jedinečná každému mluvčímu
2 *Metrika* – nauka o rytmu a metru veršů

- Mascol, *Curves of Pauline and Pseudo-Pauline style*, 1888 [28, 29]

  Články popisují často opomíjenou [39], ve své době však bezpochyby nejkomplexnější metodu stylometrické analýzy. Přestože nebyla navržena jako univerzálně použitelná, byla výjimečná zejména tím, že se zabývala několika jazykovými jevy zároveň. Popis této metody je předmětem přílohy A.

S cílem popsat autorův styl bylo navrženo mnoho číselných ukazatelů jako jsou pravděpodobnostní rozložení délky slov (Mendenhall [31], Fucks [12]), různé míry bohatosti slovní zásoby (Yule [49], Simpson [46], Sichel [45]) a další. Žádná z nich se však sama o sobě neukázala jako dostatečně účinná [16, 23] a jejich nedostatky vedly k výzkumu komplexních, vícerozměrných metod analýzy literárního stylu. Van Halteren ukázal, že každému autorovi lze přiřadit tzv. stylom[3] – sadu měřitelných rysů a jejich hodnot, které jsou pro autora určující [19]. O jazykovém rysu tvrdíme, že je součástí autorova stylomu, pokud:

a) jeho hodnota odlišuje autora od většiny ostatních autorů [19]

b) je autor v jeho použití konzistentní [41]

c) je nezávislý na tématu [9]

3.1 Atributy dokumentů

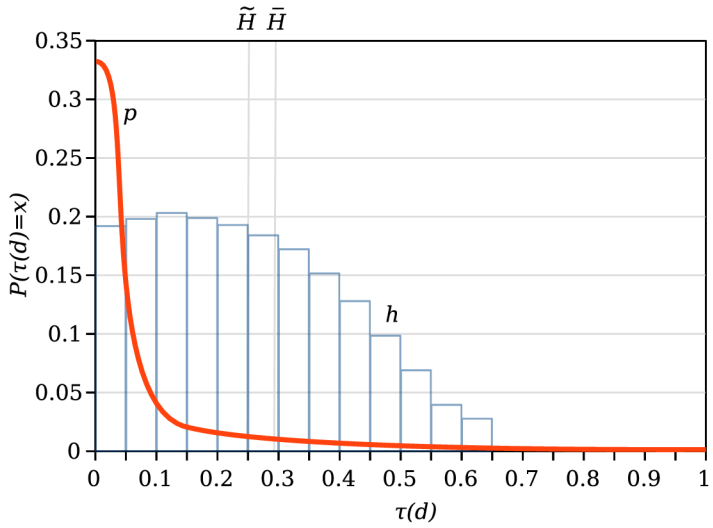
Následující stránky jsou věnovány kritériím jazykových rysů, které autorův stylom utvářejí, a způsobu, jakým jsou v rámci této práce vyhodnocovány. Při popisu používáme pojmy atribut a hodnota – atributem máme na mysli konkrétní jazykový rys, jeho hodnotou pak racionální číslo $0 \leq x \leq 1$. Interpretace této hodnoty je pro každý atribut zcela odlišná a není součástí této práce.

Dokument d z množiny všech možných dokumentů $\mathcal{D}$ nabývá pro atribut t hodnoty $\tau_t(d) = x$. Podobně pro sadu dokumentů $\mathcal{D}_a$ autora a platí $\tau_t(\mathcal{D}_a) = \{x \in \langle 0, 1\rangle \mid \exists d \in \mathcal{D}_a \ : \ \tau_t(d) = x\}$. Definujme

[3] Někdy označován také jako writeprint [25]

dále funkci $\alpha(d) = a$, která libovolnému dokumentu d přiřazuje jeho autora a z množiny všech možných autorů $\mathcal{A}$.

Atribut může napříč dokumenty jediného autora nabývat různých hodnot a pro další úvahy potřebujeme tyto hodnoty popsat. Obrázek 3.1 zřetelně ukazuje, že libovolná míra polohy jako aritmetický průměr nebo medián je pro popis stylomu nedostačující.



Obrázek 3.1: Ilustrativní srovnání hodnoty atributu dokumentů jediného autora (histogram h množiny H) a populace (křivka p). Přestože autorovy dokumenty nabývají hodnoty 0 téměř stejně často, jako jakékoli jiné hodnoty, v populaci se hodnota 0 vyskytuje s vysokou pravděpodobností. Hodnotu mediánu ($\tilde{H}$) ani aritmetického průměru ($\bar{H}$) proto nelze pro srovnání s populací použít.

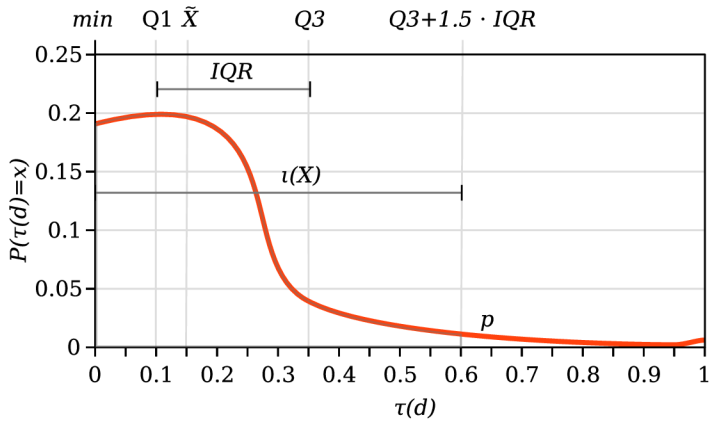
Samotný rozsah hodnot je k popisu rovněž nevhodný, protože zahrnuje také nevýznamné, odlehlé hodnoty, které mohou další výpočty nepřirozeně zkreslovat[4]. Takové výjimky jsou v jazyce zcela běžné a bez újmy na obecnosti je proto lze z výpočtů vyloučit. Jako

[4] Při automatickém určování autorství mohou nevýznamné, odlehlé hodnoty vznikat například tehdy, je-li jediný dokument v jednom nebo více atributech zcela odlišný od ostatních dokumentů sady.

efektivní rozsah $\iota(X)$ hodnot X pro unimodální rozdělení proto budeme uvažovat průnik intervalů:

$$\langle Q_1(X) - 1.5 \cdot IQR(X), Q_3(X) + 1.5 \cdot IQR(X)\rangle \cap \langle min(X), max(X)\rangle$$

Q_1 a Q_3 označují první, respektive třetí kvartil a IQR jejich absolutní rozdíl. Obrázek 3.2 ilustruje rozdíl mezi rozsahem a efektivním rozsahem.



Obrázek 3.2: Srovnání rozsahu a efektivního rozsahu hodnot. Efektivní rozsah zahrnuje všechny statisticky významné hodnoty, zatímco odlehlé hodnoty (v tomto případě $\tau(d) > 0.6$) v něm neleží.

Ne vždy je však rozložení hodnot atributu unimodální – může mít i několik vrcholů. V tom případě ztrácí definovaný efektivní rozsah na účinnosti (uvažme případ znázorněný na obrázku 3.3, ve kterém $\iota(X) = \langle 0, 1\rangle$) a je třeba uvažovat o hodnotách atributu v několika shlucích.

Navrhovaný postup rozdělení jednorozměrných dat do shluků:

(1) **Odhad křivky hustoty pravděpodobnosti**
K zběžnému odhadu rozložení lze využít histogram, ten však může být k následnému určení počtu vrcholů zavádějící – např. na obrázku 3.3 by metoda histogramu identifikovala 3 vrcholy.

Navrhujeme proto využití křivky odhadované hustoty se záměrně nízkým rozlišením. Odhadem hustoty, parametry ovlivňujícími rozlišení a konkrétní implementací v rámci této diplomové práce se zabývá kapitola 5.

(2) **Identifikace vrcholů odhadované křivky**
Lokální extrémy křivky leží v krajních bodech zkoumaného intervalu a v bodech, ve kterých směrnice její tečny mění své znaménko. Pokud se směrnice v daném bodě mění z kladné na zápornou, jde o lokální maximum. Počet vrcholů je vstupním parametrem dalšího kroku – představuje počet shluků, do kterých budou data klasifikována.

(3) **Klasifikace dat do shluků**
Aby bylo možné se shluky dále pracovat, je třeba, aby byly hodnoty v těchto skupinách co nejvíce podobné a rozdíly mezi jednotlivými shluky co největší. Tento problém řeší Jenksova optimalizační metoda [7, 1]: minimalizuje rozptyl jednotlivých tříd a maximalizuje vzájemné rozdíly mezi nimi. Podobným problémem se zabývá také kartografie[5]. Podrobný přehled metod užívaných při návrhu map a jejich objektivní i subjektivní srovnání uvádí Brewer [1].

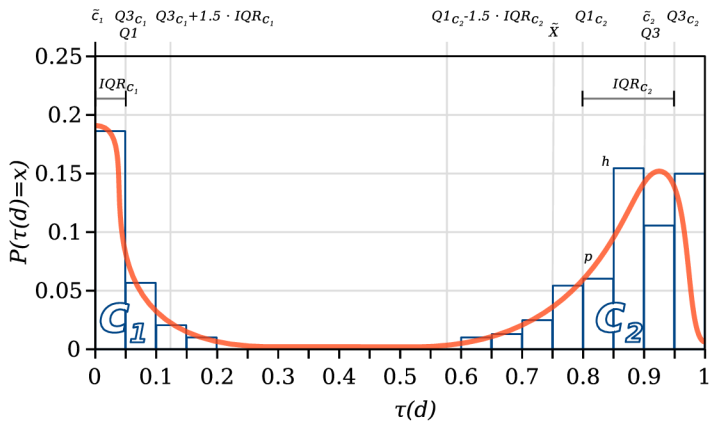
Dále pro autora $a \in \mathcal{A}$, atribut t a n shluků $c_1, \ldots, c_n \subseteq \tau_t(\mathcal{D}_a)$ definujeme efektivní interval $\mathcal{I}_a^t$:

$$\mathcal{I}_a^t = \bigcup_{1 \leq i \leq n} \iota(c_i) \tag{3.1}$$

Každému shluku c_i můžeme přiřadit váhu w_i; ta vyjadřuje poměr velikosti shluku hodnot vůči všem hodnotám, kterých autorovy dokumenty nabývají:

$$w_i = \frac{|\{x \in c_i \mid x \in \iota(c_i)\}|}{|\{x \in \tau_t(\mathcal{D}_a) \mid x \in \mathcal{I}_a^t\}|} \tag{3.2}$$

[5] Definice problému choropletových map: Za předpokladu, že má kartograf k dispozici pouze omezený počet barev k zachycení hodnoty naměřené v určité zeměpisné oblasti, jakou barvou je vhodné označit danou oblast, aby byla výsledná mapa nejpřehlednější?



Obrázek 3.3: Znázornění odhadované křivky, shluků c_1 a c_2 a jejich efektivních rozsahů

3.2 Korpus autorských textů

Vzhledem k velikosti populace a množství myslitelných textů každého člověka prakticky nelze srovnávat sporné dokumenty s dokumenty všech možných autorů (ne vždy jsou navíc texty určitého autora dostupné). Proto budeme zde uvažovaná tvrzení vztahovat ke korpusu autorských textů – tedy k souboru dokumentů všech autorů, kteří by mohli být řešením zkoumaných úloh. Popisem reprezentace a obsahu používaného korpusu se zabývá podkapitola 4.1.

Pro zpřesnění výsledků je vhodné do korpusu zahrnout pouze dokumenty, které jsou stejné délky a žánru jako zkoumaný text – tyto faktory mohou automatické určování autorství silně ovlivňovat (jak je popsáno dříve v podkapitole 1.3.1).

Množinu všech autorů dokumentů z korpusu $\mathcal{K} \subset \mathcal{D}$ budeme označovat $\mathcal{A}_\mathcal{K} \subset \mathcal{A}$:

$$\mathcal{A}_\mathcal{K} = \{a \in \mathcal{A} \mid \exists d \in \mathcal{K} \ : \ \alpha(d) = a\}$$

Při definici stylomu a atributů, které jej tvoří, uvažujeme o míře, s jakou se dokumenty autora $a \in \mathcal{A}$ v hodnotě daného atributu t liší od populace (v tomto případě od všech $b \in \mathcal{A}_\mathcal{K} \setminus \{a\}$). Vyjádřeme tuto

míru jako pravděpodobnost, že libovolný dokument $d \in \mathcal{K}$ jiného autora bude nabývat podobných hodnot, tedy:

$$P\left(\tau_t(d) \in \mathcal{I}_a^t \mid \alpha(d) \neq a\right) = \sum_{b \in \mathcal{A}_\mathcal{K} \setminus \{a\}} P\left(\tau_t(d) \in \mathcal{I}_a^t \mid \alpha(d) = b\right) =$$

$$= \sum_{b \in \mathcal{A}_\mathcal{K} \setminus \{a\}} \frac{|\{d \in \mathcal{D}_b \mid \tau_t(d) \in \mathcal{I}_a^t\}|}{|\mathcal{D}_b|} \tag{3.3}$$

$\mathcal{D}_b$: všechny dokumenty autora b
$\tau_t(d)$: hodnota atributu t pro dokument d
$\mathcal{I}_a^t$: efektivní interval, def. 3.1

Obdobně pro celou populaci $\mathcal{A}$ (jejíž je $\mathcal{A}_\mathcal{K}$ reprezentativním vzorkem) definujme pro libovolný atribut t a autora $a \in \mathcal{A}$ pravděpodobnostní funkci π_a^t:

$$\pi_a^t(x) = P(\tau_t(d) = x \mid \alpha(d) \neq a) \tag{3.4}$$

$x \in \langle 0, 1 \rangle$: libovolná hodnota atributu t
$\tau_t(d)$: hodnota atributu t pro dokument d

Její hodnotou je pravděpodobnost, že dokument jiného autora v celé populaci nabývá pro atribut t právě hodnoty x. Pokud jedinou hodnotu x nahradíme intervalem $\langle x, y \rangle$, pak:

$$P(\tau_t(d) \in \langle y, z \rangle \mid \alpha(d) \neq a) = \int_y^z \pi_a^t(x) \, \mathrm{d}x$$

Analogicky pro n shluků $c_1, \ldots, c_n \subseteq \tau_t(\mathcal{D}_a)$ a jejich váhy $w_1, \ldots, w_n$ (def. 3.2), definujeme pravděpodobnost Π_a^t:

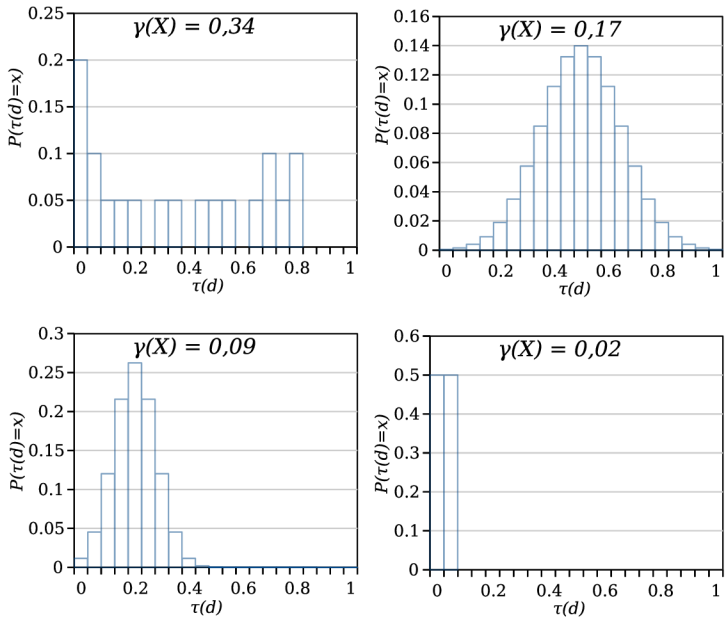
$$\Pi_a^t = P(\tau_t(d) \in \mathcal{I}_a^t \mid \alpha(d) \neq a) = \sum_{1 \leq i \leq n} w_i \int_{min(\iota(c_i))}^{max(\iota(c_i))} \pi_a^t(x) \, \mathrm{d}x \tag{3.5}$$

$\alpha(d)$: funkce přiřazující dokumentu jeho autora
$\tau_t(d)$: hodnota atributu t pro dokument d
$\iota(c_i)$: efektivní rozsah shluku c_i
$\mathcal{I}_a^t$: efektivní interval, def: 3.1

3.3 Konzistence

Autor je v použití určitého jazykového jevu konzistentní, pokud se hodnoty příslušného atributu všech jeho dokumentů navzájem příliš neliší. Jedním z primitivních ukazatelů konzistence je počet shluků, ve kterých se tyto hodnoty nacházejí. Čím více lokálních maxim nabývá odhadované rozdělení pravděpodobnosti, tím méně je autor ve svém použití atributu stálý. Samotný počet shluků však jako spolehlivá míra nestačí, protože každý z nich může být libovolně široký. Jako míru konzistence proto navrhujeme průměrný relativní rozdíl k pozorovaných hodnot $\tau_t(\mathcal{D}_a) = \{x_1, \ldots, x_k\}$ (za předpokladu unimodálního rozložení):

$$\gamma(\mathcal{D}_a) = \frac{2}{k^2} \sum_{\substack{1 \leq i < k \\ i < j \leq k}} |x_i - x_j|$$



Obrázek 3.4: Hodnota funkce γ pro různá rozdělení pravděpodobnosti. Jde o ilustrativní příklady náhodného, normálního a konstantního rozložení.

Zahrneme-li do výpočtu jen hodnoty v efektivním intervalu $\mathcal{I}_a^t$ (def. 3.1) s n shluky $c_1, \ldots, c_n \subseteq \tau_t(\mathcal{D}_a)$ a váhami $w_1, \ldots, w_n$(def. 3.2), pak míru konzistence vyjadřuje funkce Γ:

$$\Gamma(\tau_t(\mathcal{D}_a)) = \sum_{1 \leq i \leq n} w_i \, \gamma(\{x \in c_i \mid x \in \iota(c_i)\}) \tag{3.6}$$

$\iota(c_i)$: efektivní rozsah shluku c_i
$\mathcal{D}_a$: sada dokumentů autora a
$\tau_t(\mathcal{D}_a)$: hodnoty atributu t pro dokumenty autora a

Čím nižší hodnoty tento koeficient nabývá, tím jsou si jednotlivé hodnoty v efektivním rozsahu více podobné a tím více konzistentní je autor v použití daného atributu.

3.4 Rozlišovací síla atributu

Poslední vlastností atributu, kterou se budeme zabývat, je jeho rozlišovací síla pro daný typ dokumentů. Typy dokumentů, které zde uvažujeme jsou popsány v podkapitole 1.3. Každé výpočetní metodě jsme přiřadili váhu odpovídající její úspěšnosti $score(m)$ pro daný typ dokumentu. Data, která jsou podkladem k výpočtu váhy, jsou předmětem přílohy B. Nejúspěšnější metodě m_1 byla přiřazena váha 1, každé další i. nejúspěšnější metodě m_i pak podle vzorce

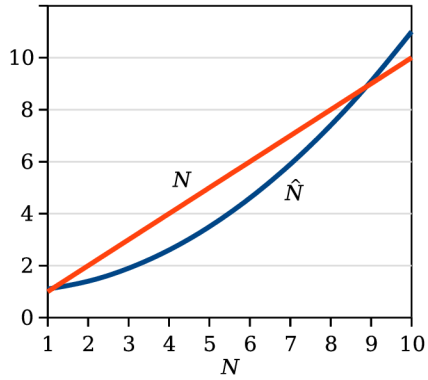
$$w(m_i) = \sqrt[3]{\frac{score(m_i)}{score(m_1)}} \tag{3.7}$$

3.5 Jedinečnost atributu

Definovali jsme nejvýznamnější faktory, které ovlivňují jedinečnost atributu a jeho hodnot pro daného autora – tedy pravděpodobnost shody s jiným autorem, stálost použití a úspěšnost výpočetní metody. Pro potřeby výpočtu koeficientu jedinečnosti byly na základě dat, naměřených v průběhu implementace, zvoleny následující škálované faktory:

- **Počet shluků**

$$\hat{\mathcal{N}}_t^a = 1 + \left(\frac{(\mathcal{N}_t^a)^2}{10}\right)$$

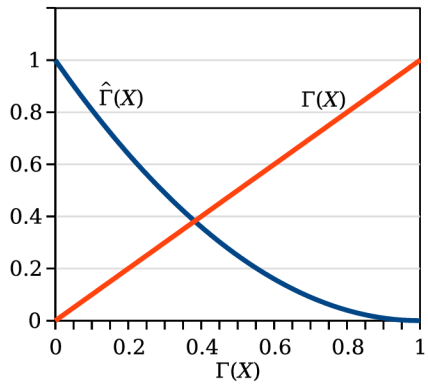


Je zřejmé, že v čím více shlucích (def. 3.1) se hodnoty atributu t jediného autora a nacházejí, tím je v užívání daného jazykového prostředku méně konzistentní a stálý. Tento faktor nabývá nízkých hodnot pro $\mathcal{N}_t^a \leq 3$, pro vyšší $\mathcal{N}_t^a$ pak jeho hodnota exponenciálně roste a následně výrazně snižuje koeficient jedinečnosti atributu.

- **Konzistence** (funkce Γ z def. 3.6)

$$\hat{\Gamma}(\tau_t(\mathcal{D}_a)) = (1 - \Gamma(\tau_t(\mathcal{D}_a)))^2$$

$\mathcal{D}_a$: Sada dokumentů autora a
$\tau_t(\mathcal{D}_a)$: Hodnoty atributu t pro sadu dokumentů $\mathcal{D}_a$

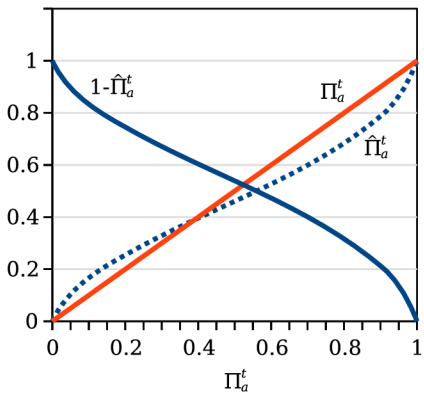


Škála tohoto faktoru zejména nadhodnocuje význam velmi vysoké konzistence a ze stylomu efektivně vylučuje atributy, pro které se $\Gamma(\tau_t(\mathcal{D}_a))$ blíží jedné – tj. v jejichž použití není autor příliš stálý.

- **Pravděpodobnost** (Π_a^t z def. 3.5)

$$\hat{\Pi}_a^t = \frac{{\Pi_a^t}^{\beta}}{\left({\Pi_a^t}^{\beta} + (1 - \Pi_a^t)^{\beta}\right)^{1/\beta}}$$

$$\overline{\beta = 0.71}$$



Jak ukazuje schéma napravo, zvolená škála výrazně zvětšuje rozdíly mezi hodnotami původní pravděpodobnosti Π_a^t v intervalu $\langle 0, 0.15 \rangle$. Proto je v tomto intervalu výsledný koeficient jedinečnosti více citlivý na odchylky pravděpodobnosti – podobně jako mnozí z nás vnímají rozdíl jednoho procentního bodu mezi 1 % a 2 % jako podstatnější, než stejný rozdíl mezi 34 % a 35 %. Psychologií subjektivního vnímání pravděpodobnosti se zabývají např. Gonzalez a Wu [14, 15]. Při návrhu škály (stejně jako při určování parametru β) jsme vycházeli právě z jejich práce.

Pokud byl atribut t vypočten pomocí metody $\mathcal{M}(t) = m$ (seznam uvažovaných metod je uveden v podkapitole 1.3.1), pak pro jeho koeficient jedinečnosti platí:

$$\Upsilon_a^t = \frac{\hat{\Pi}_a^t \cdot \hat{\Gamma}(\tau_t(\mathcal{D}_a)) \cdot w(\mathcal{M}(t))}{\hat{\mathcal{N}}}$$

$\hat{\Pi}_a^t$: faktor pravděpodobnosti, že dokumenty libovolného autora nabývají pro atribut t podobných hodnot jako dokumenty autora a.
$\hat{\Gamma}(\tau_t(\mathcal{D}_a)$: faktor konzistence hodnot atributu t napříč dokumenty $\mathcal{D}_a$ autora a.
$w(\mathcal{M}(t)$: faktor rozlišovací síly výpočetní metody $\mathcal{M}(t)$ atributu t.
$\hat{\mathcal{N}}$: koeficient počtu shluků, do kterých byly hodnoty atributu t dokumentů autora a rozděleny.

Do stylomu autora a jsou zařazeny atributy s největší hodnotou koeficientu Υ. Autorův styl pak není popsán jednotlivými atributy, ale jejich jedinečnou kombinací. Je velmi nepravděpodobné, že by se autor v použití jediného atributu natolik vymykal zbytku populace, abychom jej mohli prohlásit za určující.

Zatímco většina současných přístupů k identifikaci stylomu využívá principů strojového učení, které kladou důraz na rozlišovací schopnost stylomu jako celku nehledě na jeho velikost a komplexitu (van Helteren [19], Li [25], Rygl [41]), výše navrhované statistické pojetí stylomu vyzdvihuje význam jednotlivých kritérií, a proto je vhodnější pro úlohu automatického generování shrnutí autorova stylu.

4 Analýza řešení úkolů určování autorství

4.1 Popis systému ART

Tato práce by nebyla možná bez pokročilého systému určování autorství českých textů. Vyvíjený generátor shrnutí autorova stylu vychází z výsledků nástroje ART (*Authorship Recognition Tool*) [40], sestávajícího z několika částí (uvádíme jen seznam relevantní našemu použití):

- **Preprocesor**
  První úlohou při automatické analýze libovolného textu je jeho předzpracování – o to se v systému ART stará modul preprocesoru. Ten text zpracovává ve čtyřech krocích:

  1. Veškerá metadata jako HTML značky nebo publikační informace jsou z textu odstraněna; obrázky a jiné obsahově důležité přílohy dokumentu jsou nahrazeny jejich textovou reprezentací.
  2. Dále je text tokenizován, tj. rozdělen na jednotlivé slovní tvary (tokeny) – nejde nutně jen o slova, ale také o čísla a interpunkční znaménka.
  3. Každému slovu je jednoznačně[1] přiřazeno lemma a jeho morfologická značka. Tuto úlohu v popisovaném systému řeší morfologický analyzátor pro češtinu Majka [53], který k morfologické syntéze používá také vyvíjený generátor.
  4. Takto označkovaný text je rozdělen na samostatné věty a každá z nich je podrobena syntaktické analýze pomocí již zmiňovaného nástroje SET [24] – ten pro každou větu generuje závislostní a složkový strom.

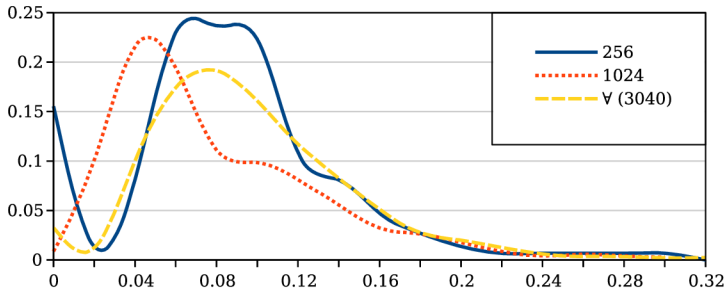
  Výsledky preprocesoru jsou podkladem pro výpočet charakteristik vyjmenovaných v podkapitole 1.3.1. Výpočet všech atributů je součástí systému ART.

[1] Správný tvar slova je zvolen procesem zvaným desambiguace

- **Korpus dokumentů**
  ART disponuje rozsáhlou relační PostgreSQL[2] databází automaticky získávaných dokumentů. Každý dokument obsahuje vyjma označkovaných slov (ve formě vertikálního textu) také výsledky syntaktické analýzy, čistý text, název, zdroj a jméno autora (jmenujeme jen vybrané vlastnosti).

  V rámci implementační části práce byl pro potřeby generátoru systém ART doplněn o možnost exportu korpusových dat ve formátu XML. Aby byla data vždy aktuální, je export dat prováděn při každém dotazu[3] – k následnému zpracování je proto nevhodné přenášet celou databázi (především kvůli její velikosti a výpočetní náročnosti exportu). Z tohoto důvodu byl navržen formát, který splňuje následující požadavky:

  a) V souladu s definicí v podkapitole 3.2 obsahuje generovaný korpus jen dokumenty stejného typu jako sporné dokumenty. Je zřejmé, že přesnost veškerých výpočtů na základě dat z korpusu je ovlivněna jeho velikostí – proto ji lze parametricky měnit. Obrázek 4.1 ukazuje vliv velikosti korpusu na odhadovanou křivku pravděpodobnosti.



Obrázek 4.1: Hustota pravděpodobnosti hodnoty atributu *k4* (číslovky) metody *Frekvence morfologických kategorií* pro různě veliké korpusy dokumentů (příspěvků v diskuzním fóru). Korpusy o velikosti 256 a 1024 jsou náhodné podmnožiny korpusu 3040 dokumentů.

[2] The PostgreSQL Global Development Group. *PostgreSQL* [online]. Dostupné z:`http://www.postgresql.org/`

[3] Tuto funkcionalitu je možné konfiguračním parametrem vypnout

b) Exportovaný korpus neobsahuje tělo (obsah) jednotlivých dokumentů, ale jen jejich číselné identifikátory a informaci o počtu různých autorů, jejichž dokumenty jsou v korpusu zahrnuty.

c) Nejpodstatnější částí generovaného souboru jsou hodnoty atributů[4] pro všechny dokumenty. Tyto jsou uspořádány tak, aby bylo možné snadno odlišit hodnoty různých autorů – předpoklad nutný pro výpočet pravděpodobnosti Π_a^t (def. 3.5).

- **Strojové učení**
  Jádrem automatického určování autorství systémem ART jsou algoritmy strojového učení. Tato část aplikace zahrnuje vícevrstvé strojové učení, které vychází z metody SVM[5] [21, 40]. Nástroj nám dovoluje srovnat autorství dvou libovolných dokumentů – tj. nejen označit shodu či rozpor, ale také vyjádřit pravděpodobnost správnosti podaného rozhodnutí.

  Výsledky těchto algoritmů (nehledě na jejich konkrétní implementaci) zpracovává vyvíjený generátor shrnutí. Pro reprezentaci výsledků byl navržen a do nástroje ART implementován export v jednotném formátu XML, který obsahuje níže uvedená data použitelná při následném zpracování:

  a) Zpracovaný obsah srovnávaných dokumentů a jejich typ

  b) Hodnoty atributů všech dokumentů

  c) Podobnosti dvojic dokumentů a rozhodnutí, zda mají stejného autora. Protože je tento krok výpočetně náročnější než kroky ostatní, jsou porovnávány jen dokumenty, které nepatří do stejné sady (sady dokumentů jsou vstupním parametrem generátoru) – vzájemná podobnost textů jediné sady z hlediska autorství nemá na výsledný výstup generátoru vliv.

---

[4] Hodnoty $x \in \langle 0, 1 \rangle$ jsou zaokrouhleny na 3 desetinná místa, ostatní jsou v exportovaném souboru vynechány.

[5] *SVM*: Support Vector Machines

4.2 Řešení vytyčených úkolů

Implementovaný generátor ve spolupráci s ostatními součástmi nástroje ART vytváří shrnutí výsledků určování autorství pro tři předeslané úlohy: verifikaci autorství, volba autora dokumentu z několika kandidátů a shlukování dokumentů podle autorství. Popis systému ART a jeho vzájemné součinnosti s generátorem je předmětem předchozí podkapitoly, zaměřme se proto nyní na řešení konkrétních problémů.

Výstupem generátoru je dokument shrnující řešení úkolů a charakteristické rysy autora (popřípadě více autorů), které toto řešení obhajují. Aby bylo shrnutí snadno interpretovatelné[6], je slovní popis výsledků daného úkolu doplněn grafickým zdůvodněním. Níže pro každou z řešených úloh uvádíme očekávané vstupní parametry, popisujeme princip jejího řešení a obsah shrnutí. Příkladem vygenerovaného dokumentu je zvláštní příloha *Zpráva o výsledcích automatického určení autorství*.

4.2.1 Verifikace autorství

Vstup generátoru:

Právě dvě sady dokumentů $\mathcal{D}_a$ a $\mathcal{D}_b$, o nichž má být rozhodnuto, zda mají stejného autora, tedy zda $a \stackrel{?}{=} b$.

$a, b \in \mathcal{A}$, množina všech možných autorů
$\mathcal{D}_a, \mathcal{D}_b \subset \mathcal{D}$, množina všech dokumentů

Princip řešení

Potvrzující rozhodnutí, tedy že autorem dokumentů v obou sadách je tentýž člověk, budeme označovat $a \simeq b$. Opačné rozhodnutí, tedy že vstupní sady dokumentů mají různé autory, značíme $a \not\simeq b$. K rozhodování využívá generátor výsledků systému ART exportovaných ve formátu XML.

Implementovaný postup řešení a jeho teoretická východiska:

[6] Interpretací numerických výsledků se zabýváme v podkapitole 1.3.3.

1. ART pomocí technik strojového učení nejprve určí míru podobnosti všech dvojic dokumentů z různých sad; tuto budeme dále označovat $\sigma(d_1, d_2)$, kde $d_1 \in \mathcal{D}_a$ a $d_2 \in \mathcal{D}_b$. Pokud je hodnota podobnosti $\sigma(d_1, d_2) > 0.5$, pak ART označí autory těchto dokumentů za shodné, tj. $\alpha(d_1) \simeq \alpha(d_2)$.

2. Tento krok budeme nazývat hlasování. Máme $n = |\mathcal{D}_a| \cdot |\mathcal{D}_b|$ hlasů, z toho kladných

$$v^+ = \sum_{(d_1,d_2)\in\mathcal{D}_a\times\mathcal{D}_b} \begin{cases} 1 & : \sigma(d_1, d_2) > 0.5 \\ 0 & : \sigma(d_1, d_2) \leq 0.5 \end{cases}$$

Pro záporné hlasy pak platí $v^- = \overline{n - v^+}$

Pokud $v^+ > \frac{n}{2}$, pak tvrdíme, že sady dokumentů mají stejného autora, tj. $\alpha(\mathcal{D}_a) \simeq \alpha(\mathcal{D}_b)$, tedy $a \simeq b$. Při hlasování můžeme dále uvažovat míru jistoty, se kterou lze řešení považovat za správné.

Definujme nejprve pravděpodobnost $\mathcal{P}_k^n$, že se alespoň k hlasujících rozhodlo správně, tj.

$$\mathcal{P}_k^n = \begin{cases} P(v^+ \geq k \mid a = b) & : a \simeq b \\ P(v^- \geq k \mid a \neq b) & : a \not\simeq b \end{cases}$$

Podobně také pravděpodobnost, že se alespoň k hlasujících zmýlilo:

$$\hat{\mathcal{P}}_k^n = \begin{cases} P(v^+ \geq k \mid a \neq b) & : a \simeq b \\ P(v^- \geq k \mid a = b) & : a \not\simeq b \end{cases} \tag{4.1}$$

Pravděpodobností v různých hlasovacích systémech se již v 18. století zabýval francouzský matematik markýz Nicolas de Condorcet. Ve svém díle *Essai sur I'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* [6] se věnoval především otázce správnosti rozhodnutí ve většinovém hlasování. Výsledkem jeho výzkumu byla následující tvrzení,:

Condorcet Jury Theorem [6, 35]:

$$\mathcal{P}_k^n = \sum_{i=k}^{n} \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} \tag{4.2}$$

p: kompetence hlasujícího, tj. pravděpodobnost správného rozhodnutí. (Condorcet předpokládá, že hlasující tvoří z hlediska kompetence homogenní skupinu.)

Dále pokud $\frac{1}{2} < p < 1$ a $k > \frac{n}{2}$, pak hodnota $\mathcal{P}_k^n$ s rostoucím n monotónně roste a v závislosti na kompetenci p je od určitého n správný výsledek garantován:

$$\lim_{n\to\infty} \mathcal{P}_k^n = 1$$

Analogicky pokud $0 < p < \frac{1}{2}$ a $k > \frac{n}{2}$, pak hodnota $\mathcal{P}_k^n$ s rostoucím n monotónně klesá a platí:

$$\lim_{n\to\infty} \mathcal{P}_k^n = 0$$

Vycházíme-li ze závěrů Condorcetova teorému, pak z Bayesovy věty plyne, že:

$$P(a = b \mid v^+ \geq k) = \frac{P(v^+ \geq k \mid a = b) \cdot P(a = b)}{P(v^+ \geq k)} =$$

$$= \frac{\mathcal{P}_k^n \cdot P(a = b)}{\mathcal{P}_k^n \cdot P(a = b) + \hat{\mathcal{P}}_k^n \cdot P(a \neq b)}$$

Obdobný vzorec pro výpočet míry jistoty užívá také systém ART
Stejně jako Rygl budeme uvažovat $P(a = b) = P(a \neq b) = 0.5$
$\mathcal{P}_k^n$: pravděpodobnost, že nejméně k hlasů bylo správných, def. 4.2
$\hat{\mathcal{P}}_k^n$: pravděpodobnost, že nejméně k hlasů bylo chybných, def. 4.1

Tuto pravděpodobnost budeme označovat jako míru jistoty v rozhodnutí, že autoři a a b jsou shodní, tj. $a \simeq b$, jestliže se pro shodu našlo alespoň k hlasů. Obdobně lze definovat také míru jistoty pro opačné rozhodnutí.

Condorcetův předpoklad homogenity hlasujících se v praxi ukázal jako nereálný, bylo proto navrženo několik zobecnění původního teorému. Mějme n hlasujících, pak $p_1, \ldots, p_n$ jsou kompetence každého z nich; v oblasti verifikace autorství budeme za kompetenci p považovat podobnost dvou dokumentů $d_1 \in \mathcal{D}_a$ a $d_2 \in \mathcal{D}_b$, tedy $\sigma(d_1, d_2)$.

Condorcet Jury Theorem zobecněný pro symetricky rozložené kompetence (Grofman [17, s. 268]):

Jsou-li kompetence p_i rozloženy symetricky, je možné bez újmy zaměnit v definici Condorcetova teorému (def. 4.2) parametr p průměrem $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$.

Toto zobecnění jsme při implementaci uvažovali, nepodařilo se nám však prokázat, že by podobnost dokumentů vycházela ze symetrického rozložení (zejména proto, že hodnoty mnohých atributů nejsou symetricky rozložené a podobnost vychází z jejich vzájemného rozdílu [7]).

Condorcet Jury Theorem zobecněný pro libovolné rozložení kompetencí (Owen [35, s. 5, def. (A1)]):

Nechť $\mathcal{N} = (1, \ldots, n)$. Jsou-li kompetence $p_1, \ldots, p_n$ vzájemně nezávislé, pak pravděpodobnost, že většina (značená jako k) rozhodla správně je rovna

$$\mathcal{F}(p_1, \ldots, p_n) = \sum_{\substack{\mathcal{S} \subset \mathcal{N} \\ |\mathcal{S}| \geq k}} \prod_{i \in \mathcal{S}} p_i \cdot \prod_{i \notin \mathcal{S}} (1 - p_i) \tag{4.3}$$

Triviálně pokud pro $\forall i \in \mathcal{N} : p_i = p$, pak

$$\mathcal{F}(p, \ldots, p) = \sum_{i=k}^{n} \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} \quad = \mathcal{P}_k^n \text{ (def. 4.2)}$$

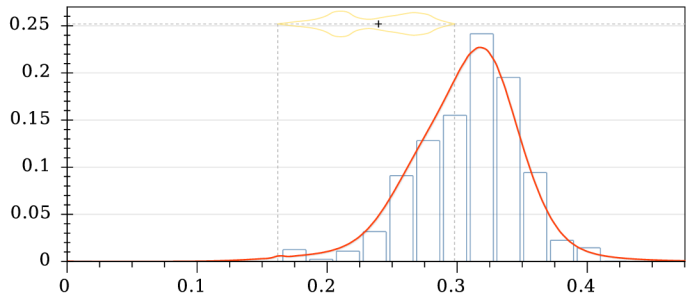
Tuto variantu jsme do generátoru implementovali, při větším počtu dokumentů se však ukázala být kvůli své výpočetní složitosti prakticky nepoužitelná.

[7] Překvapivé naopak bylo zjištění, že hodnoty některých atributů vycházejí z normálního rozložení (např. frekvence výskytu podstatných jmen ve větě)

Výstup v přirozeném jazyce:

Generovaný dokument obsahuje řešení zadané úlohy a míru jistoty, se kterou můžeme toto řešení považovat za správné.

A. Pokud $a \simeq b$, pak výstup obsahuje také n atributů, které tvoří autorův stylom (definován v kap. 3). Vhodnost výběru atributu do stylomu je, mimo jiné, zdůvodněna několika proloženými grafy (příkladem takového grafu je obrázek 4.2).



Obrázek 4.2: Příklad generovaného grafu pro atribut t: `k1-kI` metody bigramů slovních druhů, který vysvětluje zařazení atributu do stylomu autora a.

- Histogram znázorňuje hustotu pravděpodobnosti

  $$P(\tau_t(d) \in (intStart, intEnd\rangle \mid \alpha(d) \neq a)$$

  Tedy pravděpodobnosti, že jiný autor (jehož dokumenty jsou obsaženy v korpusu) nabývá hodnot v interalu $(intStart, intEnd\rangle$, tj. intervalu, který zahrnuje každý sloupec histogramu.
- Liniový graf představuje odhadovanou křivku hustoty pravděpodobnosti $\pi_a^t(x)$ z def. 3.4.
- Houslový graf pak reprezentuje rozložení hodnot atributu (hodnoty z intervalu $\mathcal{I}_a^t$, def. 3.1) ve zkoumaných dokumentech. Křížek ve středu houslového grafu označuje medián daného shluku.

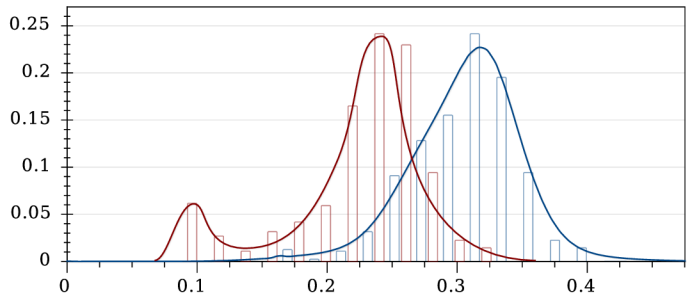
B. Pokud $a \not\simeq b$, pak výstup obsahuje n atributů, kterými se autoři nejvíce liší. Jako míru odlišnosti atributů jsme zvolili průměrný rozdíl Δ histogramů jejich rozložení [2].

Nechť srovnávané histogramy popisují rozložení náhodných proměnných X, Y, které představují hodnoty daného atributu, a $h_{X,1}, \ldots, h_{X,n}$ jsou sloupce histogramu rozložení X.

Výšku sloupce $h_{X,i}$ definujeme jako $\eta(h_{X,i})$, pak

$$\Delta(X, Y)) = \sum_{i=1}^{n} \frac{|h_{X,i} - h_{Y,i}|}{n}$$

Tento případ ilustruje obrázek 4.3.



Obrázek 4.3: Příklad generovaného srovnání rozdílných hodnot atributu.

4.2.2 Přiřazení autora

Vstup generátoru:

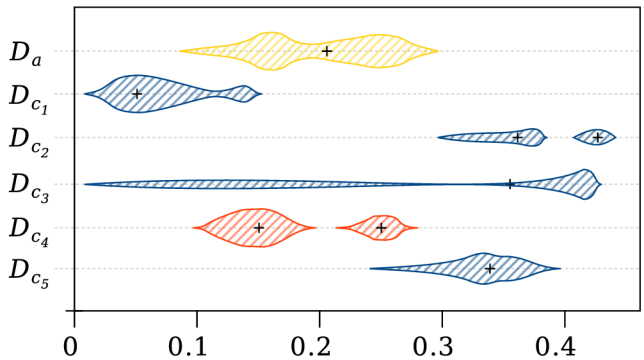
Sady dokumentů $\mathcal{D}_a, \mathcal{D}_{c_1}, \ldots, \mathcal{D}_{c_n}$, kde $n \geq 1$. Sada $\mathcal{D}_a$ obsahuje dokumenty neznámého autora a; sady $\mathcal{D}_{c_1}, \ldots, \mathcal{D}_{c_n}$ obsahují dokumenty kandidátů $c_1, \ldots, c_n$.

Princip řešení:

Proces shodný s předchozím úkolem je aplikován na všechny dvojice $(\mathcal{D}_a, \mathcal{D}_{c_i})$. Za autora sady $\mathcal{D}_a$ je označen ten kandidát c_i, jehož míra jistoty shody s a je nejvyšší. Pokud ani u jediného z kandidátů c_i nebyla nalezena shoda s a, za autora není označen žádný z nich.

Výstup v přirozeném jazyce:

Výstupní dokument obsahuje rozhodnutí, který z kandidátních autorů byl zvolen jako řešení úlohy a popisuje jeho stylom. Ve shrnutí jsou uvedeny také atributy, kterými se zvolený autor nejvíce liší od ostatních kandidátů. Srovnání uvádí obrázek 4.4.



Obrázek 4.4: Srovnání hodnot atributu t: `kl-kI` metody bigramů slovních druhů pro sady dokumentů $\mathcal{D}_a, \mathcal{D}_{c_1}, \ldots, \mathcal{D}_{c_5}$. Za autora byl zvolen $a \simeq c_3$.

4.2.3 Shlukování dle autorství

Vstup generátoru:

Sady dokumentů $\mathcal{D}_{c_1}, \ldots, \mathcal{D}_{c_n}$, každá z nich představuje shluk dokumentů určený shlukovou analýzou.

Princip řešení:

Obdobný princip jako při hledání odlišných atributů při verifikaci autora použijeme k nalezení atributů, kterými se shluky od sebe navzájem nejvíce liší a zároveň jsou pro ně jedinečné.

Výstup v přirozeném jazyce:

Generované shrnutí obsahuje seznam atributů, které shluky navzájem nejvíce odlišují.

5 Implementace generátoru shrnutí

Poslední kapitola této diplomové práce se zabývá samotnou implementací nástroje pro automatické shrnutí výstupu algoritmů určování autorství v přirozeném jazyce. Protože jsme v předešlých kapitolách uvažovali především určování autorství českých textů (stejně jako používaný systém ART), je generované shrnutí rovněž v českém jazyce.

5.1 Jádrový odhad hustoty (Kernel Density Estimation)

Při odhadu hustoty pravděpodobnosti v aplikaci využíváme metodu jádrového odhadu hustoty. Ve stručnosti uvádíme teoretické základy, na kterých tato technika staví; podrobně ji pak popisuje například Řezáč [52].

Jádrovou funkcí (zkráceně jádro) nazýváme reálnou nezápornou symetrickou funkci K, pro kterou $\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$. V implementaci používáme jen Gaussovo jádro (vycházíme z předpokladu, že hodnoty atributů pocházejí ze smíšeného normálního rozložení), pro tuto metodu však bývají používána také jádra jiná:

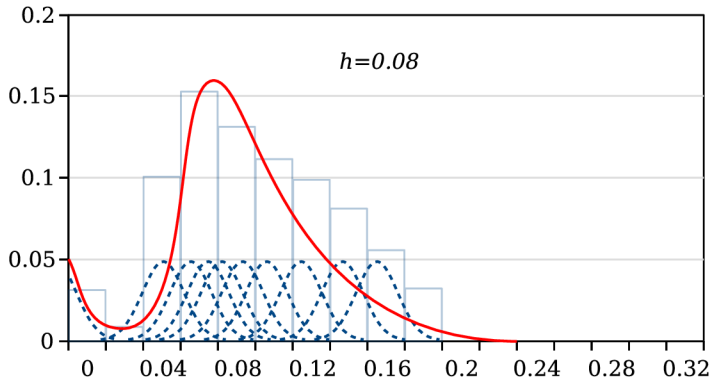
Název jádra	K(x)
Gaussovo	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$
Epanečnikovo	$\frac{3}{4}(1-x)^2$
Kvartické	$\frac{15}{16}(1-x^2)^2$
Trojúhelníkové	$1-\lvert x\rvert$

Obrázek 5.1: Příklady jader používaných při metodě jádrového odhadu hustoty pravděpodobnosti

Nechť $x_1, \ldots, x_n$ jsou nezávislé realizace náhodné proměnné $\mathcal{X}$ s neznámou hustotou pravděpodobnosti f. Odhadovanou hustotu pravděpodobnosti značíme $\hat{f}$ a definujeme vztahem:

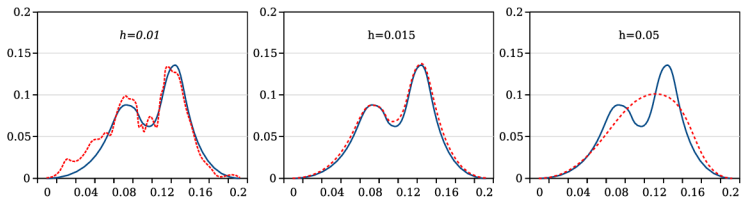
$$\hat{f}_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{5.1}$$

h: šířka vyhlazovacího okna, tzv. vyhlazovací parametr



Obrázek 5.2: Ilustrace principu metody Kernel Density Estimation.Přerušovaně jsou vyznačena Gaussova jádra, proložená všemi pozorovanými body. Výsledná křivka je dána součtem hodnot jader v daném bodě.

Šířka vyhlazovacího okna a volba jádrové funkce jsou faktory, které nejvíce ovlivňují tvar odhadované křivky. Obrázek 5.3 ukazuje vliv hodnoty vyhlazovacího parametru na odhadovanou křivku.



Obrázek 5.3: Srovnání vlivu vyhlazovacího parametru h na odhadovanou křivku. Plná křivka znázorňuje hustotu f, přerušovaná pak $\hat{f}$.

V generátoru jsme implementovali několik metod[1] pro automatickou volbu parametru h na základě dat $x_1, \ldots, x_n$. Jak však poznamenává Minnotte [33], v případech, kdy nejsou data unimodálně rozložena, je obtížné nalézt takovou hodnotu parametru h, která by adekvátně odlišovala jednotlivé lokální extrémy a jak jsme již popsali dříve, hodnoty atributů vycházejí z unimodálního rozložení jen zřídkakdy. Algoritmus jádrového odhadu hustoty pravděpodobnosti jsme proto rozšířili o adaptivní vyhlazovací parametr h_i, jehož hodnota se mění v závislosti na hodnotě x_i.

Parametr h_i v bodě x_i je dán vztahem [42]:

$$h_i = h \left(\frac{\hat{f}(x_i)}{g} \right)^{-0.5}$$

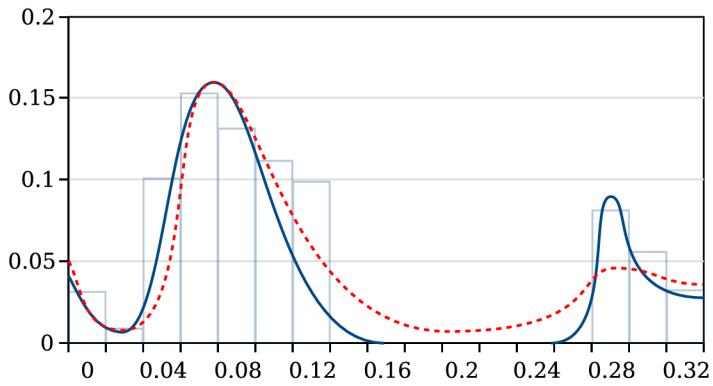
$\hat{f}(x_i)$: hodnota odhadované hustoty pravděpodobnosti s fixním h
g: geometrický průměr hodnot $x_1, \ldots, x_n$

Pro odhadovanou funkci hustoty pravděpodobnosti pak platí

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} \cdot K \left(\frac{x - x_i}{h_i} \right)$$

n: počet pozorování
K: jádrová funkce (v našem případě Gaussova standardní hustota)

[1] Používanou metodu lze měnit v konfiguraci generátoru.



Obrázek 5.4: Srovnání odhadované křivky s fixním h (přerušovaně) a adaptivním h. Jak lze vidět na grafu, fixní parametr – přestože věrohodně zachytí rozložení v intervalu $\langle 0, 0.14 \rangle$ – příliš vyhladí hodnoty v intervalu $\langle 0.16, 0.34 \rangle$

5.2 Použité softwarové nástroje

- **ART** [40]
  Generátor vychází z dříve zmiňovaného systému ART; ten je vytvořen v programovacím jazyce Python a sám využívá různých nástrojů pro strojové učení a zpracování přirozeného jazyka.

- **Java, Python**
  Implementovaná aplikace je vytvořena v jazyce Java. Pro potřeby přenosu výsledků určování autorství bylo třeba rozšířit systém ART o exportní skripty v jazyce Python.

- **Majka** [53]
  Při automatické tvorbě výstupu v přirozeném jazyce používá generátor k morfologické syntéze některých výrazů dříve uváděný analyzátor českého jazyka Majka.

- **LaTeX**
  Shrnutí je generováno z konfigurovatelné šablony do souboru formátu LaTeX. Ten je následně vysázen do dokumentu PDF programem, uvedeným v konfiguračním souboru.

- **Externí knihovny**
  Pro snadnější práci s programovými parametry a XML soubory je program závislý na knihovnách args4j[2] a dom4j[3].

  Součástí shrnutí jsou také grafy; ty jsou generovány s pomocí knihoven Gral[4] a VectorGraphics2D[5].

- **Převzaté algoritmy**
  Při implementaci jsme převzali následující algoritmy:

  - **Integrace standardní Gaussovy funkce**
    Při výpočtu pravděpodobností počítáme s plochou pod křivkou hustoty pravděpodobnosti normálního rozložení. Algoritmus výpočtu integrálu jsme převzali z knihovny Colt[6].

  - **Jenksova optimalizační metoda**
    V kapitole 3 o autorově stylomu jsme se odkazovali na Jenksovu metodu, kterou používáme ke klasifikaci jednorozměrných dat do shluků. Při implementaci jsme vycházeli z částečně chybného algoritmu v JavaScriptu[7] a původního algoritmu ve Fortranu[8].

---

2 KAWAGUCHI, Koshuke. *args4j: v. 2.0.21* [online, MIT License, cit. 2014-05-22]. Dostupné z: `http://args4j.kohsuke.org/`.

3 *dom4j: v.2.0* [online, BSD License, cit. 2014-05-22]. Dostupné z: `http://dom4j.sourceforge.net/`.

4 SEIFERT, E; SEIFERT, M. *GRAL* [online, LGPL v3, cit. 2014-05-22]. Dostupné z: `http://trac.erichseifert.de/gral/`.

5 SEIFERT, E; SEIFERT, M. *VectorGraphics2D* [online, LGPL v3, cit. 2014-05.22]. Dostupné z: `http://trac.erichseifert.de/vectorgraphics2d`.

6 CERN - European Organization for Nuclear Research. *Colt* [online, cit. 2014-05-22]. Dostupné z: `https://github.com/carlsonp/Colt/tree/master/src/cern/jet`.

7 MACWRIGHT, Tom. *Simple Statistics* [online, cit. 2014-05-22]. Dostupné z: `http://www.macwright.org/simple-statistics/`.

8 *CMLIB* [online, cit. 2014-05-22]. 1988. Dostupné z: `http://lib.stat.cmu.edu/cmlib/src/cluster/fish.f`

6 Závěr

Práce shrnuje základy problému určování autorství, uvádí historický přehled užívaných technik a příklady jejich použití. Navzdory nepřípustnosti výsledků automatického určování autorství v českém soudnictví považujeme zkoumanou oblast za důležitou, zejména pro její přesah do mnoha praktických oborů.

Cílem této práce bylo také analyzovat užití určování autorství v soudnictví, současný stav výzkumu charakteristických rysů autora, jeho stylomu a používaných metod určování autorství. Z této analýzy jsme vycházeli při návrhu statistických metod identifikace stylomu a následné implementaci.

Klíčovým přínosem této diplomové práce je systematický popis stylomu a integrace navrhovaných metod do nástroje určování autorství ART. Implementovaný generátor vytváří snadno interpretovatelný dokument shrnující řešení vytyčených úloh v přirozeném jazyce. Při generování využíváme nástrojů zpracování přirozeného jazyka, např. morfologický analyzátor češtiny Majka. Jako příklad výstupu pro úlohu verifikace autorství uvádíme zvláštní přílohu *Zpráva o výsledcích automatického určení autorství*.

Bibliografie

1 BREWER, Cynthia; PICKLE, Linda. Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series. *Annals of the Association of American Geographers*. 2002, roč. 92, č. 4, s. 662–681. Dostupný také z WWW: ⟨http://dx.doi.org/10.1111/1467-8306.00310⟩.

2 CAO, Yang; PETZOLD, Linda. Accuracy limitations and the measurement of errors in the stochastic simulation of chemically reacting systems. *Journal of Computational Physics*. 2006, roč. 212, č. 1, s. 6–24. Dostupný také z WWW: ⟨http://dx.doi.org/10.1016/j.jcp.2005.06.012⟩. ISSN 0021-9991.

3 *Ceglia v. Zuckerberg et al.* New York : New York Western District Court, 2010. Online [cit. 2014-05-22]. Dostupné z:⟨http://dockets.justia.com/docket/new-york/nywdce/1:2010cv00569/798⟩.

4 CHASKI, Carol E. The Computational-Linguistic Approach to Forensic Authorship Attribution. In. *Law and Language: Theory and Practice*. Düsseldorf : Düsseldorf University Press, 2006. Dostupný také z WWW: ⟨http://alinguisticevidence.org/wp-content/uploads/2012/07/Chaski_2006CompLingAuthorIDChapter1.pdf⟩.

5 *Commonwealth vs. Richard Wayne Illes, Sr.* 2004. Online [cit. 2014-05-22]. Dostupné z:⟨http://www.lycolaw.org/Cases/opinions/2005/Illes070805br.pdf⟩.

6 CONDORCET, Nicolas. *Essai sur I'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paříž : Imprimerie Royale, 1785. Dostupný také z WWW: ⟨https://archive.org/details/essaisurlapplica00cond⟩.

7 COULSON, M. R. C. In The Matter Of Class Intervals For Choropleth Maps : With Particular Reference To The Work Of George F Jenks. *Cartographica*. 1987, roč. 24, č. 2.

8 CRYSTAL, David; DAVY, Derek. *Investigating English style*. Londýn : Longman, 1969. ISBN 0-582-52212-9.

9 DAELMANS, Walter. Explanation in Computational Stylometry. In. *CICLing (2)*. 2013, s. 451–462. Dostupný také z WWW: ⟨http://dx.doi.org/10.1007/978-3-642-37256-8_37⟩.

10 DOBROVSKÝ, Josef. *Fragmentum Pragense evangelii S. Marci, vulgo autographi*. 1778.

11 FLEAY, F. G. *Shakespeare manual*. Londýn : Macmillan a Co., 1876.

12 FUCKS, Wilhelm. On Mathematical Analysis of Style. *Biometrika*. 1952, roč. 39, č. 1–2, s. 122–129. Dostupný také z WWW: ⟨http://dx.doi.org/10.1093/biomet/39.1-2.122⟩.

13 GEBAUER, Jan. *Poučení o padělaných rukopisích Královedvorském a Zelenohorském : pro širší kruhy inteligence české*. Praha : L. Masaryk v Hustopeči, 1888. Dostupný také z WWW: ⟨http://kramerius.nkp.cz/kramerius/handle/ABA001/1147162⟩.

14 GONZALEZ, R.; WU, G. Curvature of the probability weighting function. *Management Science*. Roč. 42, č. 12, s. 1676–1690.

15 GONZALEZ, R.; WU, G. On the Shape of the Probability Weighting Function. *Cognitive Psychology*. 1999, roč. 38, č. 1, s. 129–166.

16 GRIEVE, Jack William. *Quantitative Authorship Attribution : A History and an Evaluation of Techniques*. Burnaby : Simon Fraser University, 2005.

17 GROFMAN, Bernard; OWEN, Guillermo; FELD, Scott L. Thirteen theorems in search of the truth. *Theory and Decision*. 1983, roč. 15, č. 3, s. 261–278. Dostupný také z WWW: ⟨http://dx.doi.org/10.1007/BF00125672⟩. ISSN 0040-5833.

18 GÖNÜL, M. Sinan; ÖNKAL; LAWRENCE. The effects of structural characteristics of explanations on use of a DSS. *Decision Support Systems*. 2006, roč. 42, č. 3, s. 1481–1493. Dostupný také z WWW: ⟨http://dx.doi.org/10.1016/j.dss.2005.12.003⟩.

19 HALTEREN, Hans van et al. New Machine Learning Methods Demonstrate the Existence of a Human Stylome. *Journal of Quantitative Linguistics*. 2005, roč. 12, č. 1, s. 65–77. Dostupný také z WWW: ⟨http://dx.doi.org/10.1080/09296170500055350⟩.

20 HOLLINGSWORTH, Charles D. *Syntactic Stylometry : Using Sentence Structure for Authorship Attribution*. 2012.

21 HSU, Chih-Wei; CHANG, Chih-Chung; LIN, Chih-Jen. *A Practical Guide to Support Vector Classification*. Taipei, Taiwan : Department of Computer Science, National Taiwan University, 2003. Aktualizováno 2010. Dostupný také z WWW: ⟨http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf⟩.

22 JUOLA, Patrick. Authorship Attribution. *Foundations and Trends® in Information Retrieval*. 2008, roč. 1, č. 3, s. 233–334. Dostupný také z WWW: ⟨http://dx.doi.org/10.1561/1500000005⟩. ISSN 1554-0669.

23 KOPPEL, Moshe; SCHLER, Jonathan; ARGAMON, Shlomo. Computational methods in authorship attribution. *Journal of the Association for Information Science and Technology*. 2009, roč. 60, č. 1, s. 9–26. Dostupný také z WWW: ⟨http://dx.doi.org/10.1002/asi.20961⟩.

24 KOVÁŘ, Vojtěch; HORÁK, Aleš; JAKUBÍČEK, Miloš. Syntactic Analysis Using Finite Patterns: A New Parsing System for Czech. In. *Human Language Technology. Challenges for Computer Science and Linguistics*. Berlin/Heidelberg : Springer, 2011, s. 161–171.

25 LI, Jiexun; ZHENG, Rong; CHEN, Hsinchun. From fingerprint to writeprint. *Communications of the ACM*. 2006, roč. 49, s. 82.

26 LOVE, Harold. *Attributing Authorship : An Introduction*. New York : Cambridge University Press, 2002. ISBN 9781139146586.

27 LUYCKX, Kim; DAELEMANS, Walter. Authorship Attribution and Verification with Many Authors and Limited Data. In. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester : Coling 2008 Organizing Committee, s. 513–520.

28 MASCOL, Conrad. Curves of Pauline and Pseudo-Pauline style I. *The Unitarian Review*. 1888, roč. 30, č. 5, s. 452–460. Dostupný také z WWW: ⟨https://archive.org/stream/unitarianreview00allegoog⟩.

29 MASCOL, Conrad. Curves of Pauline and Pseudo-Pauline style II. *The Unitarian Review*. 1888, roč. 30, č. 6, s. 539–546. Dostupný také z WWW: ⟨https://archive.org/stream/unitarianreview00allegoog⟩.

30 MCMENAMIN, Gerald R. *Forensic Linguistics : Advances in Forensic Stylistics*. New York : CRC Press, 2002. ISBN 9781420041170.

31 MENDENHALL, T. C. The Characteristic Curves of Composition. *Science*. 1887, roč. 9, č. 214, s. 237–249. Dostupný také z WWW: ⟨http://www.jstor.org/stable/1764604⟩. ISSN 00368075.

32 MENDENHALL, T. C. A Mechanical solution of a Literary Problem. *The Popular Science Montly*. 1901, č. 60, s. 97–105. Dostupný také z WWW: ⟨http://www.archive.org/details/popularsciencemo60newy⟩.

33 MINNOTTE, Michael C. *A test of mode existence with applications to multimodality*. 1993.

34 MITCHELL, Tom. *Machine Learning*. New York : McGraw-Hill, 1997. ISBN 978-0070428072.

35 OWEN, Guillermo; GROFMAN, Bernard; FELD, Scott L. Proving a distribution-free generalization of the Condorcet Jury Theorem. *Mathematical Social Sciences*. 1989, roč. 17, č. 1, s. 1–16. Dostupný také z WWW: ⟨http://dx.doi.org/10.1016/0165-4896(89)90012-7⟩. ISSN 0165-4896.

36 PORTET, François; REITER, Ehud; GATT, Albert et al. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*. 2009, roč. 173, č. 7–8, s. 789–816. Dostupný také z WWW: ⟨http://dx.doi.org/10.1016/j.artint.2008.12.002⟩.

37 PUPPE, Frank; ATZMUELLER, Martin; BUSCHER, Georg et al. Application and Evaluation of a Medical Knowledge System in Sonography (SONOCONSULT). In. *Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence*. Amsterdam : IOS Press, 2008, s. 683–687. Dostupný také z WWW: ⟨http://dx.doi.org/10.3233/978-1-58603-891-5-683⟩. ISBN 978-1-58603-891-5.

38 REITER, Ehud et al. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*. 2005, roč. 167, č. 1–2, s. 137–169. Dostupný také z WWW: ⟨http://dx.doi.org/10.1016/j.artint.2005.06.006⟩.

39 RUDMAN, Joseph. The State of Authorship Attribution Studies : Some Problems and Solutions. *Computers and the Humanities*. 1997, roč. 31, č. 4, s. 351–365. Dostupný také z WWW: ⟨http://dx.doi.org/10.1023/A:1001018624850⟩.

40 RYGL, Jan. *Determining Authorship of Anonymous Texts*. Rigorózní práce. Brno : Masarykova Univerzita, 2013.

41 RYGL, Jan. Automatic adaptation of author's stylometric features to document types. In. *Text, Speech and Dialogue - 17th International Conference*. Brno : Springer, 2014 [in press].

42 SAIN, Stephan R. *Adaptive Kernel Density Estimation*. Houston, Texas : Rice University, 1994.

43 SEDLÁČEK, Radek. *Ajka tagset*. 2006. Online [cit. 2014-05-19]. Dostupné z:⟨http://nlp.fi.muni.cz/projekty/ajka/tags.pdf⟩.

44 SHERMAN, L. A. Some Observations upon the Sentence-Length in English Prose. *University Studies of University of Nebraska*. 1888, roč. I, č. II, s. 119–130. Dostupný také z WWW: ⟨https://archive.org/stream/universitystudie1188892univ⟩.

45 SICHEL, H. S. On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association*. 1975, roč. 70, č. 351, s. 542–547.

46 SIMPSON, Edward H. Measurement of Diversity. *Nature*. 1949, č. 163, s. 688.

47 SOLAN, Lawrence M. Intuition versus Algorithm : The Case of Forensic Authorship Attribution. *Brooklyn Journal of Law and Policy*. 2013, roč. 21, č. 551. Dostupný také z WWW: ⟨http://ssrn.com/abstract=2272090⟩.

48 YU, Guoxing. Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics*. 2010, roč. 31, č. 2, s. 236–259.

49 YULE, Udny G. *The Statistical Study of Literary Vocabulary*. Cambridge : Cambridge University Press, 1944.

50 ZHANG, XueYing. Route Description Using Natural Language Generation Technology. In LI, Hang; LIU, Ting; MA, Wei-Ying et al. (ed.). *Information Retrieval Technology*. 2008, s. 454–459. Lecture Notes in Computer Science. Dostupný také z WWW: ⟨http://dx.doi.org/10.1007/978-3-540-68636-1_47⟩.

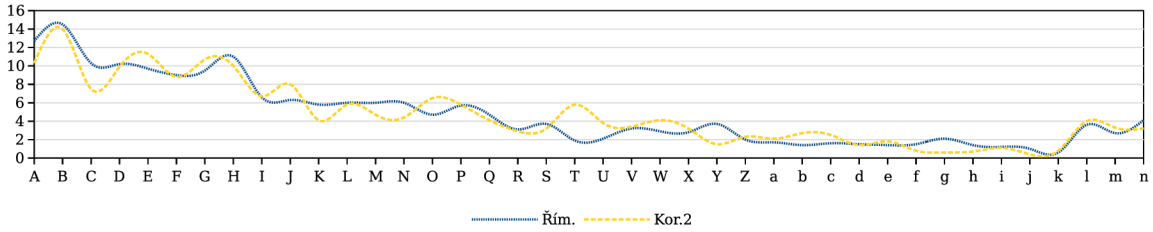
51 Zákon č. 121 ze dne 7. dubna 2000 o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon). In. *Sbírka zákonů, Česká republika*. 2000. ISSN: 1211-1244.

52 ŘEZÁČ, Martin. *Jádrové odhady hustoty*. Disertační práce. Brno : Masarykova Unvierzita, 2007.

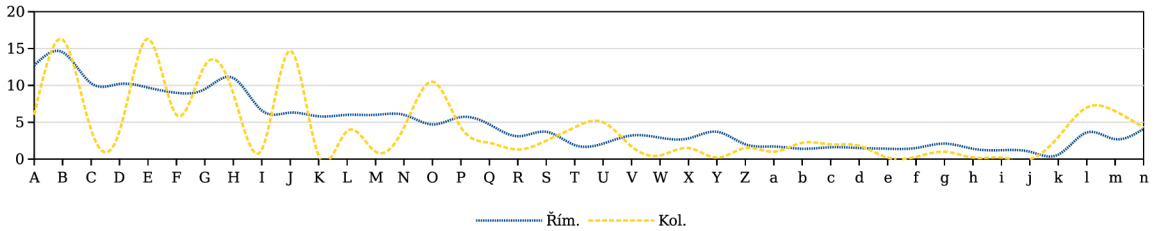
53 ŠMERK, Pavel. *K počítačové morfologické analýze češtiny*. Disertační práce. Brno : Masarykova univerzita, Fakulta informatiky, 2010. Dostupný také z WWW: ⟨http://is.muni.cz/th/3880/fi_d/⟩.

Přílohy

A. *Curves of style* listů sv. Pavla



Obrázek A.1: *Curves of style* listů Římanům a 2. Korintským



Obrázek A.2: *Curves of style* listů Římanům a Koloským

Obrázky 1 a 2 jsou rekonstrukcí křivek (původní křivky nebyly z technických důvodů publikovány), která vychází z tabulek uvedených v závěru Mascolova pojednání *Curves of Pauline and Pseudo-Pauline style*. Obrázek 1 vykresluje křivku dvou epištol, o nichž autor tvrdí, že mají stejného autora, křivka na obrázku 2 pak srovnává listy, o jejichž shodném autorství Mascol pochybuje. Křivky sledují 40 různých atributů jako počet vět(A), teček(F), infinitivů(Z) a dalších jazykových znaků.[28, 29]

B. Úspěšnost charakteristik pro různé typy dokumentů

	TD_1		TD_2		TD_3		TD_4		TD_5	
metoda	$score$	w	$score$	w	$score$	w	$score$	w	$score$	w
Průměrná délka vět	11.07	**0.88**	16.08	**0.92**	11.89	**0.84**	12.65	**0.90**	13.35	**0.88**
Frekvence morf. značek	10.27	**0.85**	10.01	**0.79**	13.21	**0.87**	10.95	**0.86**	11.22	**0.83**
Frekv. morf. kategorií	5.87	**0.71**	8.81	**0.76**	17.01	**0.94**	8.54	**0.79**	6.92	**0.71**
Frekv. slovních druhů	16.46	**1.00**	12.33	**0.84**	7.74	**0.73**	13.08	**0.91**	9.34	**0.78**
Frekv. bigramů slovních druhů	5.23	**0.68**	11.99	**0.84**	11.78	**0.84**	14.71	**0.95**	11.25	**0.83**
Frekvence stop slov	4.35	**0.64**	20.45	**1.00**	11.36	**0.83**	6.76	**0.73**	16.17	**0.94**
Opakování slov	16.07	**0.99**	5.83	**0.66**	14.81	**0.90**	7.87	**0.77**	15.62	**0.93**
Syntaktická analýza	8.15	**0.79**	9.17	**0.77**	9.76	**0.79**	4.27	**0.63**	8.2	**0.75**
Syntaktická analýza n-gramů	9.27	**0.83**	10.93	**0.81**	11.43	**0.83**	7.71	**0.77**	10.91	**0.82**
Bohatost slovní zásoby	3.95	**0.62**	9.48	**0.77**	10.21	**0.80**	11.21	**0.87**	14.7	**0.91**
Velká písmena	10.77	**0.87**	13.4	**0.87**	7.1	**0.71**	5.8	**0.70**	19.55	**1.00**
Frekv. interpunkčních znamének	9.37	**0.83**	12.87	**0.86**	13.3	**0.87**	16.71	**0.99**	12.31	**0.86**
Typografické chyby	10.22	**0.85**	15.91	**0.92**	20.16	**1.00**	17.2	**1.00**	11.42	**0.84**
Užití emotikon	11.53	**0.89**	9.25	**0.77**	13.72	**0.88**	14.09	**0.94**	12.61	**0.86**

Typy dokumentů:

TD_1: Knihy
TD_2: Články na blogu
TD_3: Příspěvky v diskuzním fóru
TD_4: Komentáře k internetovým článkům
TD_5: Tweety

$score$: Účinnost metody pro daný typ českého dokumentu, jak uvádí Rygl[41]
w: Váha metody dle def. 3.7