Filozofická fakulta Univerzity Palackého v Olomouci
Katedra obecné lingvistiky

Identifikace autora ve forenzní lingvistice

bakalářská diplomová práce

Autor: Anna Tichá

Vedoucí práce: Mgr. Lukáš Zámečník, Ph.D.

Olomouc 2019

Prohlašuji, že jsem diplomovou práci na téma „Identifikace autora ve forenzní lingvistice“ vypracovala samostatně a s použitím uvedené literatury a pramenů.

V Olomouci dne 20. 8. 2019 Podpis:

Poděkování

Chtěla bych poděkovat Mgr. Lukáši Zámečníkovi, PhD., za vedení mé bakalářské práce, cenné rady a odborný dohled. Děkuji také Mgr. Vladimíru Matlachovi za pomoc při práci s lingvistickým softwarem QUITA a další cenné rady.

Abstrakt

1. **Příjmení a jméno autora:** Tichá Anna
2. **Název katedry a fakulty:** Katedra obecné lingvistiky, Filozofická fakulta
3. **Název bakalářské práce:** Identifikace autora ve forenzní lingvistice
4. **Jméno vedoucího bakalářské práce:** Mgr. Lukáš Zámečník, PhD.
5. **Počet znaků:** 72 939
6. **Počet příloh:** 2
7. **Počet titulů použité literatury:** 21
8. **Klíčová slova:** forenzní lingvistika, identifikace autora, kvantitativní a kvalitativní analýza, hierarchické shlukování, vícerozměrné škálování, QUITA

Cílem bakalářské práce je obeznámit čtenáře s interdisciplinárním oborem forenzní lingvistika. Práce je rozdělena na dvě části - teoretickou a praktickou. Teoretická část seznamuje čtenáře s nejvýznamnějšími osobnostmi, díly a také případy, ve kterých byla forenzní lingvistika využita v soudní praxi. Dále vymezuje, jaké druhy textů se používají k forenzně-lingvistické analýze, jak se dělí a jaké podmínky musí splňovat, aby mohly být použity k analýze. V praktické části je představena v současnosti využívaná kvalitativní analýza textů a následně jsou navrženy způsoby, jak dosavadní využívanou metodiku rozšířit za využití vícerozměrných metod. Jednotlivé metody byly demonstrovány na předem vyhotoveném korpusu textů vhodných k forenzně-lingvistické analýze.

Nejprve jsme provedli kvantitativně-kvalitativní analýzu, dále popsali vícerozměrné škálování a hierarchické shlukování. Poté jsme se zaměřili na analýzu pomocí bag-of-words modelu a metodu hodnocení podle vybraných kvantitativně lingvistických indexů. Vzhledem k výsledkům analýz je patrné, že bag-of-words model je z použitých metod nejspolehlivější. Můžeme to přičítat tomu, že texty používané k forenzně-lingvistické analýze jsou psané spontánně a projevují se v nich vlastnosti idiolektu jednotlivého pisatele. Každý člověk totiž používá zaběhlé fráze, které jsou typické pro jeho osobní projev.

Abstract

The aim of this thesis is to introduce to the readers the interdisciplinary field of forensic linguistics. The thesis is divided into two parts - theoretical and practical. The theoretical part introduces the most important persons, works and also cases in which forensic linguistics was used in forensic practice. It also defines what types of texts are used in forensic-linguistic analysis, how they are divided and what conditions they must meet to be used for analysis. The practical part presents the currently used qualitative analysis of texts and subsequently suggests ways to extend the existing methodology using multidimensional methods. Individual methods were demonstrated on a pre-made corpus of texts suitable for forensic-linguistic analysis. We first carried out a quantitative-qualitative analysis, then we introduced multidimensional scaling and hierarchical clustering. The reader was acquainted with the analysis using the bag of words model and with the method of evaluation according to selected quantitative linguistic indices. Given the results of the analyzes it is apparent that the bag-of-words model is the most successful of the methods used. This can be attributed to the fact that the texts used for forensic-linguistic analysis are written spontaneously and show the idiolect properties of the individual writer.

Úvod 8
I. Teoretická část 11
1. Co je to forenzní lingvistika 12
2. Historie a vývoj evropské forenzní lingvistiky 13
2.1. Významné osobnosti světové forenzní lingvistiky 14
2.2. Vražda Jenny Nicholl 14
3. Historie a vývoj forenzní lingvistiky v českém prostředí 17
3.1. Osobnosti a literatura české forenzní lingvistiky 17
3.2. Sumarizace 18
4. Texty ve forenzní lingvistice 19
4.1. Dělení textů podle délky 19
4.2. Sporný a srovnávací materiál 20
4.3. Sumarizace 20
II. Praktická část 21
5. Korpus textů 22
6. Kvalitativně-kvantitativní metoda 23
6.1. Kritéria hodnocení 23
6.2. Sumarizace 35
6.3. Vícerozměrné škálování (multidimensional scaling) 35
6.3.1. Závěr 38
6.3.2. Sumarizace 39
6.4. Hierarchické shlukování 39
6.4.1. Závěr 42
6.4.2. Sumarizace 42
6.5. Závěr testování hypotéz 43
7. Množina slov (bag-of-words model) 45
7.1. Závěr 48
8. Indexy 49
8.1. Závěr 55
8.2. Sumarizace 55
9. Závěr praktické části 56
Závěr 58

Seznam použité literatury 63
Přílohy..... 66

Úvod

Tématem této bakalářské práce je identifikace autora ve forenzní lingvistice. Forenzní lingvistika je interdisciplinární obor na pomezí lingvistiky a kriminalistiky, kterým se v České republice zabývá jen malé množství lidí.

Práce se skládá ze dvou částí – teoretické a praktické. Teoretická část je pojata jako seznámení s forenzní lingvistikou jako samostatným oborem a je rozdělena do čtyř kapitol. Praktická část se zabývá především analýzou textů a popisem metod, které v naší práci používáme k identifikaci autorů, a je rozdělena do pěti kapitol.

První kapitola teoretické části práce bude zaměřena na obeznámení čtenáře s oborem forenzní lingvistika. Ve druhé a třetí kapitole se budeme věnovat historii nejdříve světové, potom české forenzní lingvistiky a představíme si nejvýznamnější osoby, díla a také případy, ve kterých byla forenzní lingvistika využita v soudní praxi. V poslední kapitole teoretické části se zaměříme na to, jaké druhy textů se ve forenzní lingvistice používají a jaké podmínky musí splňovat, aby mohly být použity k analýze.

Cílem praktické části bude vytvořit stručný a přehledný souhrn metod, které se v českém prostředí používají k identifikaci autora, a následně je prakticky aplikovat na námi vybrané anonymní texty. Korpus textů popíšeme v 5. kapitole.

Budeme používat teoretické (konceptuální analýza, kompilace, srovnání aj.) i praktické metody. Jako první bude provedena kvantitativně-kvalitativní analýza, v níž si stanovíme kritéria hodnocení anonymních textů. Tato kritéria budou detailně popsána v podkapitole 6.1. Podle těchto kritérií pak na základě kvalifikovaného úsudku provedeme hodnocení jednotlivých textů na škále od 0 do 10, 10 znamená nejvíc a nula úplně chybí. Za účelem zjednodušení práce s kritérii bude vytvořeno několik tabulek. V této části se pokusíme rozvinout dosavadní používané metody tak, aby dokázaly najednou pojmout veškerá použitá kritéria a nabídnout tak analytikovi ucelený pohled na texty. V této práci tedy teprve prozkoumáváme možnosti, jak doposud využívanou metodiku rozšířit. Budeme se snažit nalézt co nejefektivnější způsob využití vícerozměrných metod (MDS, HClust), kdy je jejich využití čistě experimentální, stejně tak i použití euklidovské vzdálenosti a kosinové nepodobnosti. Tyto metody budou detailně popsány v podkapitole 6.3.

Zároveň si stanovíme několik hypotéz a na základě výsledků hodnocení se pokusíme určit, které texty by mohly patřit ke stejným autorům. Analytická část práce bude koncipována tak, aby při průzkumu a hodnocení metod autor tohoto textu neznal odpověď na otázku, které anonymní texty patří ke kterému autorovi. Toho bude docíleno tak, že texty budou anonymizovány jinou osobou. Kompletní tabulka textů a jejich autorů bude uvedena v přílohách. Veškeré testy metod a jejich efektivity budou tedy nejprve prováděny naslepo a explikují tak veškerou nejistotu.

V této kapitole budeme dále hodnotit, která kritéria se pro naši analýzu ukážou jako přínosná a která nám naopak neposkytnou informace užitečné k rozlišení jednotlivých autorů. Budeme využívat lingvistického softwaru QUITA. Vícerozměrným škálováním vytvoříme náhledy na podobnost anonymních textů a pokusíme se identifikovat shluky jednotlivých textů podle toho, zda by mohly patřit ke stejným autorům. Vícerozměrným škálováním dále vytvoříme náhledy na podobnost neanonymních textů a vyhodnotíme, zda se naše předchozí odhady ohledně autorství textů potvrdily.

V podkapitole 6.4. popíšeme metodu hierarchického shlukování a jeho grafickou interpretaci ve formě tzv. dendrogramu (stromového diagramu). Hierarchickým shlukováním vytvoříme náhledy na podobnost anonymních textů a pokusíme se identifikovat shluky textů na jednotlivých větvích podle toho, které by mohly patřit ke stejným autorům. Dále vytvoříme náhledy na podobnost neanonymních textů a vyhodnotíme, zda se naše předchozí odhady ohledně autorství potvrdily.

Po interpretaci výsledků vícerozměrného škálování a hierarchického shlukování provedeme hodnocení, zda se nám podařilo potvrdit nebo vyvrátit hypotézy, které jsme vyslovili v podkapitole 6.5.

Další metodou, kterou v práci použijeme, bude bag-of-words model, pomocí níž provedeme analýzu anonymních i neanonymních textů. Na konci této kapitoly vyhodnotíme, zda byla tato metoda úspěšná v rozlišení jednotlivých autorů.

V 8. kapitole se zaměříme na metodu hodnocení podle vybraných kvantitativně lingvistických indexů. Nejprve stručně popíšeme jednotlivé indexy, poté provedeme analýzu anonymních i neanonymních textů a následně se pokusíme vyhodnotit, zda byla tato metoda úspěšná v rozlišení jednotlivých autorů.

V závěru bakalářské práce shrneme získané poznatky a zhodnotíme, která metoda určování autorství anonymních textů se ukázala být nejúspěšnější.

I. Teoretická část

Teoretická část práce bude zaměřena na problematiku forenzní lingvistiky. Nejprve se bude věnovat historii a vývoji světové forenzní lingvistiky. Následně bude popsán jeden z nedávných případů, ve kterém sehrála forenzní lingvistika důležitou roli při odsouzení muže obviněného z vraždy. Dále se bude zabývat historií, osobnostmi a literaturou forenzní lingvistiky v českém prostředí (u nás známé spíše pod pojmem jazyková expertiza). V poslední podkapitole teoretické části stručně popíšeme, jaké texty se využívají k forenzní analýze, jak se dělí a jaké podmínky musí splňovat, aby analýza proběhla úspěšně.

1. Co je to forenzní lingvistika

Jako forenzní lingvistika se označuje jeden z nejmladších oborů aplikované lingvistiky. Jedná se o obor interdisciplinární, který se pohybuje na pomezí jazykovědy a právních věd, zejména pak kriminalistiky. Pro tento obor jsou však často využívány i poznatky ze sociologie, psychologie nebo grafologie. Mezi podobory forenzní lingvistiky spadají například forenzní stylistika, forenzní fonetika, kontrastivní lingvistika či forenzní sémantika.

V současnosti probíhají forenzně-lingvistické výzkumy ve dvou hlavních oblastech.

První oblast je zaměřena na zkoumání veškerých psaných i mluvených právních textů z hlediska jejich přesnosti, srozumitelnosti a jednoznačnosti při interpretaci, ale i ze strategického hlediska, za účelem vítězství v soudním sporu, napsání přesvědčivého odvolacího protokolu, ale třeba i sestavení legislativy či formulování znění právních dokumentů.

Druhá oblast se zabývá identifikací osoby, takzvaným profilováním autora, na základě jeho jazykového chování. Při profilování autora se forenzní lingvista pokouší co nejvíce zúžit okruh potenciálních autorů na základě vytipovaných jazykových charakteristik. [1]

Obor forenzní lingvistiky se v posledních letech velmi rychle rozvíjí, v českém akademickém prostředí však stále není příliš známou disciplínou.

[1]SVOBODOVÁ, Marie. Forenzní lingvistika: obsah a možnosti. *Slovo a slovesnost* [online]. 1997, **58**(2), 124-129 [cit. 2019-08-19]. Dostupné z: http://sas.ujc.cas.cz/archiv.php?lang=en&art=3726

2. Historie a vývoj evropské forenzní lingvistiky

Počátky forenzní lingvistiky není lehké datovat. Lidé se už od dávných dob pokoušeli o určování autorství uměleckých textů, zvláště pokud se jednalo o podezření z plagiátorství. Mezi známé případy patří například snahy o určení autora u posvátných textů nebo Shakespearových her. Termín forenzní lingvistika poprvé použil v roce 1968 profesor lingvistiky Jan Svartvik při analýze čtyř výpovědí Timothy Evanse, ve kterých se přiznal k vraždě své manželky a dcery, za což byl následně pověšen. Svartvikova zjištění spolu s dalšími důkazy prokázala, že Evans nemohl nadiktovat prohlášení, která mu byla přisuzována.

V anglickém právu byla po dlouhou dobu stanovena řada pravidel týkajících se výslechu svědků, zejména toho, jak zapisovat výpovědi. Tyto předpisy byly jednoduše označeny jako soudcovská pravidla (Judge's Rules), která stanovovala, že podezřelí mají své výpovědi policistům diktovat, policisté nemají podezřelé osoby přerušovat a že s výjimkou menších objasnění by neměly v průběhu vypovídání být podezřelým kladeny žádné otázky.

V praxi se to však téměř nikdy nedělo. Soudci, kteří formulovali pravidla pro zapisování výpovědí, si neuvědomovali, že získat doslovný přepis diktovaného prohlášení je téměř nemožné – lidé zkrátka nemluví plynule, přeskakují, vynechávají detaily a pak se k nim vrací. Policista tedy obvykle v průběhu výslechu pokládal podezřelému řadu otázek, zapisoval si pouze poznámky a pak dodatečně sepsal prohlášení podezřelého. Nepoužil však jeho původní slova, ale fráze a slovní spojení, na které byli policisté dlouhodobě zvyklí. Policejní prohlášení tedy obsahovala fráze typu „I thenobserved" atd. Toto slovní spojení však není mezi lidmi běžně používáno, spíše odráží způsob formulace, která se nazývá „policejní registr"[2] a jež je sama oblastí studia forenzní lingvistiky.

Forenzní lingvistika se postupně zařadila mezi uznávané forenzní metody, na začátku 90. let 20. století se konala série seminářů, které v roce 1993 vedly k založení International Association of Forensic Linguists (IAFL), International Association for Forensic Phonetics (IAFP) a o rok později časopisu Forensic Linguistics: The International Journal of Speech, Language and the Law. Koncem

[2]OLSSON, John. *WhatisForensicLinguistics?* [online]. , 4-5 [cit. 2019-07-02]. Dostupné z: https://www.thetext.co.uk/what_is.pdf

devadesátých let byla založena webová stránka www.iafl.org, která obsahuje mnoho užitečných zdrojů a odkazů pro členy IAFL i veřejnost.[3]

International Association of Forensic Linguists je organizace, která sdružuje lingvisty zabývající se jazykem práva. Cílem IAFL je, mimo jiné, zlepšit správu právních systémů na celém světě prostřednictvím lepšího porozumění interakci mezi jazykem a zákonem, propagace používání jazyka jako důkazu v občanskoprávních věcech (ochranná známka, smluvní spory, hanobení, odpovědnost za výrobek, podvodné obchodní praktiky, porušování autorských práv), šíření znalostí o jazykové analýze a jejích forenzních aplikacích mezi příslušnými odborníky po celém světě a v neposlední řadě shromažďování materiálů, jako jsou přiznání, sebevražedné poznámky a policejní výslechy a jejich zpřístupnění v online korpusu.[4]

Roku 1994 založil John Olsson ve Velké Británii Institut forenzní lingvistiky.[5] V současnosti je možné získat magisterský titul v tomto oboru minimálně na třech evropských univerzitách – v Cardiffu, Birminghamu a Barceloně.

2.1. Významné osobnosti světové forenzní lingvistiky

Mezi významné osobnosti zabývající se forenzní lingvistikou patří například Malcolm Coulthard, Jan Svartvik, John Olsson, Janet Cotterill a Hannes Kniffka.

2.2. Vražda Jenny Nicholl

V únoru 2008 byl za vraždu devatenáctileté Jenny Nicholl na doživotí odsouzen pětačtyřicetiletý David Hodgson – i přes to, že se k vraždě nepřiznal a tělo nebylo nikdy nalezeno. Velkou roli v jeho usvědčení hrála právě forenzní lingvistika.

Jenny Nicholl žila se svojí rodinou v Richmondu v Anglii, kde pracovala v supermarketu a hrála v rockové kapele. Už od svých čtrnácti let se vídala s Davidem Hodgsonem, který byl otcem jejích spolužaček.[6]

---

[3]BLACKWELL, Susan. History of ForensicLinguistics. *EncyclopediaofAppliedLinguistics* [online]. 2013, 1-2 [cit. 2019-07-02].
[4]*International AssociationofForensicLinguistics* [online]. [cit. 2019-07-02]. Dostupné z: https://www.iafl.org/about-iafl/
[5]http://www.thetext.co.uk/

Dne 30. června 2005 řekla Jenny rodičům, že nebude přes noc doma a sbalila si vybavení na kempování. Jenny o sobě nedala rodičům vědět několik dní, proto její zmizení nahlásili na policii (4. července 2005). O devět dní později policie vyslechla Davida Hodgsona, který popřel, že by s Jenny měl vztah a že by měl něco společného s jejím zmizením. Hned další den byl Jennyin mobil zapnut a byly z něj odeslány zprávy adresované jejímu otci a přátelům, které měly navodit dojem, že je Jenny naživu a v pořádku. Policisté však pojali podezření, že zprávy nenapsala Jenny, a k vyšetřování přizvali profesora Malcolma Coultharda, prvního profesora forenzní lingvistiky na světě,[7] který se už dříve podílel na řešení více než dvou set případů.

Policisté poskytli profesoru Coulthardovi přepisy tří textových zpráv, které byly odeslány z Jennyina mobilního telefonu v době jejího zmizení, více než sto zpráv napsaných Davidem Hodgsonem a jedenáct zpráv, které prokazatelně napsala Jenny. Zprávy na první pohled vypadaly, že mohly být napsány devatenáctiletou dívkou, ale při bližším zkoumání našel profesor Coulthard podstatné nesrovnalosti. Například Jenny i odesílatel sporných zpráv používali "2" jako náhradu za "to", jenže Jenny nenechávala mezeru mezi "2" a následujícím slovem, zatímco odesílatel ano. Dalšími očividnými rozdíly bylo, že zatímco Jenny Nicholl obvykle psávala "Im" a "Im not“, autor sporných SMS zpráv použil "I am" a "aint". Jinými příklady jsou rozdíly v tom, že Jenny Nicholl psávala "my", "cu" a "fone", ve sporných SMS zprávách se objevilo "me", "cya" a "phone".[8]

SMS od Jenny:
Shit isit.fuck **icant2day** iveallready **booked2go** bowling.cantrealypulloutwil **go2sho p** and get her sumet soon.**thanx4tdlin** me

No **im** outwiv jak sorryittookme so long **ive** had **fone** offcoz **havnt** got much battery

---

[6]*Jenny 'murdered by marriedlover'* [online]. 15. 1. 2008 [cit. 2019-07-01]. Dostupné z: http://news.bbc.co.uk/2/hi/uk_news/england/north_yorkshire/7189805.stm
[7]*ProfessorMalcolmCoulthard* [online]. [cit. 2019-07-02]. Dostupné z: https://www2.aston.ac.uk/lss/staff-directory/coulthardm
[8]OWEN, Amos. The text trap. *TheNorthern Echo* [online]. 27.2.2008 [cit. 2019-07-01]. Dostupné z: https://www.thenorthernecho.co.uk/news/2076811.the-text-trap/

Sporné SMS:

Shegotme in thisshitits her fault not mine getblame 4evrything.**i am** sorry ok just had **2** lve shes a bitch no food in alwayssearching **me** roomeating **me** sweets.**ave2** go ok **i am** very sorry **x**

Hi jen tel jak **i am** ok knowever 1 s gona b madtellthem **i am** sorry.living in scotlandwiv my boyfriend.shitting **meself** dadsgonakillmemumdontgive a **shite**.hope nik didntgrassmeup.keeping **phone** of.telldad car jumpsoutofgear and stallsputitback in auction.tellhim **i am** sorry

Zdroj SMS: HARDAKER, Claire. *The Case of Jenny Nicholl* [online]. [cit. 2019-07-01]. Dostupné z: https://wp.lancs.ac.uk/drclaireh/2012/10/01/the-case-of-jenny-nicholl/

Profesor Coulthard řekl: "*From a linguistic point of view, what I couldn't say was he sent those text messages. But what I could say was he shared a lot of the same features and wasamong a small number of possiblesenders.*" [9][10]

Díky údajům společnosti O2 se podařilo prokázat, že SMS zprávy z Jennyina telefonu byly odeslány 9. a 14. července z oblastí Brampton a Jedburgh na hranicích se Skotskem. Dalším podstatným důkazem v tomto procesu byly údaje poskytnuté půjčovnou aut, ve které si Hodgson v tomto časovém úseku pronajal auto. Ty prokazují, že počet najetých kilometrů přibližně odpovídají cestám do těchto oblastí. David Hodgson byl odsouzen a nyní si odpykává doživotní trest s možností propuštění nejdříve po osmnácti letech. Stále však trvá na své nevině.

[9]Howcarefulanalysisof text messageshelped police to catch a killer. *Darlington and StocktonTimes* [online]. 7.3.2008 [cit. 2019-07-01]. Dostupné z: https://www.darlingtonandstocktontimes.co.uk/news/2102472.how-careful-analysis-of-text-messages-helped-police-to-catch-a-killer/

[10]"Z lingvistického pohledu nemohu říci, že ty zprávy poslal. Co však mohu říci, je, že sdílel mnoho společných znaků (s autorem zpráv) a byl v úzkém okruhu možných odesílatelů."

3. Historie a vývoj forenzní lingvistiky v českém prostředí

Podle Musilové se v českém prostředí forenzní lingvistika rozvíjí od 50. let 20. století. Znalci v Kriminalistickém ústavu se věnovali písmoznalectví, ale zároveň již při sestavování posudků využívali *„některá nápadná jazyková hlediska, především užitou slovní zásobu, pravopis a interpunkci“*[11]

V současné době se forenzní lingvisté zabývají především autorstvím sporných textů, k čemuž využívají postup profilování a identifikace. V české republice mohou činnost forenzního lingvisty vykonávat jedině jmenovaní znalci. Jejich činnost se řídí zákonem č. 36/1967 Sb., o znalcích a tlumočnících, v aktuálním znění a vyhláškou č. 37/1967 Sb., k provedení zákona o znalcích a tlumočnících, v aktuálním znění. Písmoznalectvím se dle údajů Ministerstva spravedlnosti České republiky platných k 16. 08. 2019 zabývá 28 znalců a jazykovou expertizou dva znalci.[12]

3.1. Osobnosti a literatura české forenzní lingvistiky

Jednou z nejvýznamnějších osobností české forenzní lingvistiky je PhDr. Václava Musilová, soudní znalkyně, zabývající se jazykovou expertizou a písmoznalectvím. Doktorka Musilová se specializuje na určováním druhu, modelu a značky psacího stroje použitého k vyhotovení strojopisu, expertizou ručního písma a dále expertizou pravosti platidel a cenin a technickou expertizou písemností čili zjišťováním pravosti nebo způsobu vyhotovení padělaných či pozměněných písemností.[13]

Druhou soudní znalkyní v oboru jazyková expertiza je doc. PhDr. Alena Aigner, CSc., která působí na Pedagogické fakultě Jihočeské univerzity v Českých Budějovicích.

---

[11]MUSILOVÁ, Václava. Forenzní lingvistika I. *Čeština doma a ve světě*. Praha: Ústav českého jazyka a teorie komunikace FF UK, 2005a, roč. 13, 1 - 2, 66.
[12]*Ministerstvo spravedlnosti České republiky: Evidence znalců a tlumočníků* [online]. [cit. 2019-08-17]. Dostupné z: http://datalot.justice.cz/justice/repznatl.nsf/$$SearchForm?OpenForm
[13]*Soudní znalci z oboru kriminalistiky* [online]. [cit. 2019-07-01]. Dostupné z: https://www.grafickeexpertizy.com/

V osmdesátých letech minulého století se určování autorství literárních textů věnoval také doc. PhDr. Pavel Vašák, DrSc., literární teoretik a textolog, který o této metodě napsal publikaci *Metody určování autorství*.

Bohužel není dostupné větší množství česky psané odborné literatury. Většinou se jedná o články v kriminalistických a jazykovědných časopisech, například Čeština *doma a ve světě*, *Československá kriminalistika* nebo *Kriminalistický sborník*.

3.2. Sumarizace

Cílem 2. kapitoly bylo obeznámení s definicí forenzní lingvistiky, její historií ve světě i v českém prostředí. Představili jsme významné osobnosti forenzní lingvistiky a jejich zaměření a uvedli jsme literaturu pojednávající o forenzní lingvistice.

4. Texty ve forenzní lingvistice

Ke své práci forenzní lingvisté využívají dva typy materiálu: sporný a srovnávací. Aby mohl být text podroben forenzně-lingvistickému zkoumání, musí splňovat několik základních podmínek. Mezi tyto podmínky patří například souvislost textu nebo dostatečná délka textu. Podle Musilové[14] bývají forenznímu lingvistovi k expertize předkládány nejčastěji anonymní dopisy. Tyto dopisy se dělí do několika skupin podle obsahu: nejčastěji anonymní dopisy výhružné a vyděračské, dále dopisy pomlouvačné, kompromitující, obscénní, udavačské, urážlivé a dopisy upozorňující na trestnou činnost. Dále se znalci zabývají dopisy sebevrahů, deníkovými záznamy, texty administrativními a právními, u nichž se zkoumá např. protiprávní zasahování do znění písemností, smluv nebo jejich stylizování, texty administrativními, odbornými a publicistickými, závěťmi, autorstvím překladu nebo policejními a soudními protokoly. Zkoumá se autentičnost, protiprávní zasahování do znění písemností nebo jejich stylizování. V neposlední řadě se forenzní lingvistika vyjadřuje k víceznačnosti formulací tvrzení a vhodnosti terminologie zejména právních textů.

4.1. Dělení textů podle délky

Podle kriminalisty Jiřího Strause lze texty dělit do čtyř kategorií. Ve své knize *Kriminalistická technika* (2012) rozlišuje texty velmi krátké, které mají maximálně 170 slov, texty krátké (170 – 380 slov), texty dlouhé (380 – 750 slov) a texty velmi dlouhé, obsahující více než 750 slov.[15]

U textů do 500 slov se přistupuje ke kvalitativní analýze, jelikož by výsledky kvantitativní (statistické) analýzy byly nepřesné, delší texty je možné analyzovat kvantitativně. Tyto metody budou dále popsány v praktické části této bakalářské práce.

[14] MUSILOVÁ, Václava. Co je forenzní lingvistika I. *Čeština doma a ve světě*. Praha: Ústav českého jazyka a teorie komunikace FF UK, 2005a, roč. 13, 1-2, 69-70.
[15] STRAUS, Jiří et al. *Kriminalistická technika*. 3. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, 2012.

Jak bylo již dříve v této práci (kapitola Vražda Jenny Nicholl) ukázáno, v dnešní době elektronické komunikace stále nabývají na významu velmi krátké texty, například SMS, e-maily či příspěvky v internetových diskuzích. Komunikace na internetu se od běžného rozhovoru v mnohém liší. Jedním z hlavních rozdílů je, že online prostor zaručuje anonymitu, člověk není v internetové komunikaci tlačen časem, a tudíž si může svůj projev předem promyslet, vyprodukovat a poté ještě znovu přečíst a zkontrolovat.

4.2. Sporný a srovnávací materiál

Jako sporný materiál je označován *„jakýkoliv souvislý jazykový projev, který je v příčinné souvislosti s trestným činem a jehož autor není znám“*[16]. Tento materiál je v průběhu analýzy srovnáván s tzv. srovnávacím materiálem, jehož autor je prokazatelně znám. Podle Strause je nezbytně nutné, aby srovnávací materiál vznikl spontánně, aniž by si pisatel byl vědom toho, že text bude použit k analýze.

4.3. Sumarizace

Cílem této kapitoly bylo seznámení čtenáře s tím, jaké texty ke své práci forenzní lingvista používá. Jedná se o dva typy materiálu: sporný a srovnávací. Dále shrnuje, podle čeho se texty dělí a jaké podmínky musí texty splňovat, aby je bylo možné použít k analýze.

[16] STRAUS, Jiří et al. *Kriminalistická technika*. 3. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, 2012, 139.

II. Praktická část

Cílem praktické části bude popsat existující metody identifikace autora a následně navrhnout možnosti, jak dosavadní využívanou metodiku rozšířit co nejefektivnějším způsobem za využití vícerozměrných metod.

Budeme používat teoretické i praktické metody. Navrhneme několik hypotéz o tom, které texty by na základě námi zvolených kritérií mohly patřit k jednotlivým autorům, a následně se pokusíme za použití kvantitativních metod tyto hypotézy potvrdit nebo vyvrátit. Jednotlivá kritéria budou detailně rozebrána v kapitole 6.1. a zároveň k nim bude přiřazeno několik ukázek z korpusu textů vytvořeného pro tuto bakalářskou práci. Veškeré ukázky budou použity v autentické podobě.

5. Korpus textů

Analyzováno bude celkem dvacet textů od čtyř různých autorů, vždy pět textů o každého autora v rozsahu kolem 5 tisíc znaků. Veškeré texty se skládají z kratších, spontánně stylizovaných komentářů poskládaných za sebe, aby tvořily souvislý text. Takto utvořené texty jsou dostatečně dlouhé a zaměřené na stejná témata. Jedná se o komentáře, které autoři napsali zhruba ve stejném časovém období pod články na webové stránce www.svice.cz. Jedná se o server, na němž může kdokoliv publikovat článek na libovolné téma, ať je jakkoliv kontroverzní. Příkladem mohou být články nadepsané: *Měl by být rasizmus uzákoněn? Jsou invalidní důchodci na odstřel? Neexistuje lepší systém než kapitalismus? Jsou ženy, které jsou krásné, a ženy, které se umí namalovat*.

Cílem je vyvolat pod článkem diskuzi, v níž se každý přispěvatel může svobodně vyjádřit k danému tématu. Všechny texty jsou z kapacitních důvodů k dispozici na přiloženém CD.

6. Kvalitativně-kvantitativní metoda

Cílem této části je provést kvalitativně-kvantitativní analýzu zadaných textů. Straus tuto metodu ve své knize *Kriminalistická technika* definuje jako *„zjištění jednotlivých znaků jazykového vyjadřování podle jednotlivých jazykových rovin tak, aby bylo možno při jejich hodnocení poznat vyjadřovací zvyklosti autora, tzn., že se zjišťují znaky v rámci stylistické roviny, syntaktické, lexikální, morfologické, fonologické a pomocné (pravopis a písařské zvyklosti)“*.[17] Na základě této analýzy se pokusíme kvalifikovaným odhadem přiřadit jednotlivé texty k sobě tak, aby odpovídaly daným autorům. Pro účely kvalitativně-kvantitativní metody byly texty anonymizovány. Nejsou tedy označeny jménem autora, byl jim náhodně přidělen název Text 1 – Text 20.

Jako hodnotící kritéria jsme si zvolili čtrnáct znaků jazykového vyjadřování, které bývají pro jednotlivé autory typické. Těmito kritérii jsou interpunkce, používání velkých a malých písmen, citoslovcí, emotikonů, výskyt dvou a více teček za větou, oslovování, výskyt cizích slov, výskyt gramatických chyb, nespisovného jazyka a nadávek. Dále budeme hodnotit komplexitu vět, konfliktnost, rasismus a projevy nenávisti vůči minoritním skupinám, používání CapsLocku nebo jiného zvýraznění slov. V následující části hodnotíme kritéria na základě kvalifikovaného odhadu na škále od 0 do 10, 10 znamená nejvíc a nula úplně chybí.

V kvantitativní části analýzy budeme používat software pro kvantitativní analýzu dat – QUITA.[18]

6.1. Kritéria hodnocení

V následující části stručně představíme námi zvolená kritéria, která budeme hodnotit v jednotlivých textech.

[17]STRAUS, Jiří et al. *Kriminalistická technika*. 3. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, 2012. ISBN 978-80-7380-409-1.
[18]https://www.quitaonline.com/

- Interpunkce

  Toto kritérium hodnotí, v jaké míře autoři anonymních textů používají čárky a tečky ve větách a zda ve slovech píší háčky a čárky.

- Velká a malá písmena

  Toto kritérium hodnotí, v jaké míře autoři anonymních textů používají velká a malá písmena na začátku věty a při psaní vlastních jmen.

- Emotikony

  Toto kritérium hodnotí, v jaké míře autoři anonymních textů používají emotikony. Nerozlišujeme různé druhy emotikonů, kritérium zahrnuje všechny emotikony, usmívající se i mračící se, s nosem i bez nosu.

- Dvě a více teček za větou

  Toto kritérium hodnotí, v jaké míře autoři anonymních textů používají za větou vyšší počet teček než jednu.

- Oslovování

  Autoři anonymních textů se často ve svých komentářích navzájem oslovují jménem či přezdívkou a reagují nebo odkazují se na předchozí komentáře. Toto kritérium hodnotí, v jaké míře se v textu objeví oslovení jménem či přezdívkou.

- Cizí slova

  Toto kritérium hodnotí, v jaké míře autoři anonymních textů používají cizí slova a odborné výrazy.

- Gramatické chyby

  Do této kategorie zahrnujeme chyby v pravopise (pomíjíme očividné překlepy). Nejčastěji se jedná o chyby ve shodě podmětu s přísudkem a ve vyjmenovaných slovech.

- Nespisovnost

  Toto kritérium hodnotí, v jaké míře se autoři anonymních textů vyjadřují v souladu s kodifikovanou normou spisovného jazyka.

- Komplexita věty

  Toto kritérium hodnotí použití jednoduché nebo komplikované větné struktury.

- Konfliktnost

  Pomocí tohoto kritéria zjišťujeme, jak konfliktně působí daný text na čtenáře, zda autor používá útočné fráze a zda ve svých příspěvcích napadá ostatní účastníky diskuze.

- Nadávky

  Toto kritérium hodnotí, v jaké míře se v jednotlivých textech objevují vulgarismy.

- Rasismus a projevy nenávisti vůči minoritním skupinám

  Toto kritérium hodnotí, v jaké míře jsou určité skupiny lidí napadány kvůli jejich rase, sexuální orientaci, náboženskému či politickému přesvědčení.

- Citoslovce

  Toto kritérium hodnotí, v jaké míře je v jednotlivých textech objevují citoslovce. Citoslovce jsou neohebný slovní druh, který vyjadřuje zvuky, nálady mluvčího nebo pocity. Problémem u tohoto kritéria může být ne vždy jasný rozdíl mezi citoslovci a částicemi. Více se této problematice věnuje například Miloslav Vondráček ve svém článku *Citoslovce a částice – hranice slovního druhu*.[19]

- CapsLock nebo jiné zvýraznění slova

  V internetové komunikaci se funkce CapsLock běžně používá ke zdůraznění obsahu sdělení nebo jeho části, často bývá chápán např. jako ekvivalent křiku. Jako další způsob zdůraznění nebo zvýraznění slova používají diskutéři mezery, které vkládají mezi každý znak zdůrazňovaného slova. Příklad: n e j v ě t š í.

Jedním z kritérií, které by bylo možné do analýzy zahrnout, je používání rodových koncovek. Jako příklad uvádíme úryvek z Textu 1, ze kterého můžeme usuzovat, že autorem je žena:

> *Viděla jsem,jak to v poslední chvíli stočil tak,aby jí neublížil a do poslední chvíle tu malou vílu nadnášel,než žuchli na zem.*

[19]VONDRÁČEK, Miloslav. Citoslovce a částice — hranice slovního druhu. *Naše řeč* [online]. 1998, **81**(1), 29-37 [cit. 2019-07-20]. Dostupné z: http://nase-rec.ujc.cas.cz/archiv.php?art=7419

Jedná se však o příliš robustní vlastnost, kterou jsme se rozhodli do zvolených kritérií nezahrnout. Naší snahou bylo zvolit skupinu slabších kritérií, která mohou o autorství vypovídat až ve vzájemné kombinaci.

Pro hodnocení anonymních textů jsme vytvořili Tabulka_1. V levém sloupci jsou seřazeny sporné texty a v horním řádku jsou vypsány jednotlivá kritéria, které u textů hodnotíme.

	Interpunkce	Mal á a velká písmena	Emotiko-ny	Dvě a více teček	Oslovová-ní	Cizí slova	Gramatické chyby	Nespisovnost	Komplexita věty	Konfliktnost	Nadáv-ky	Rasismus	Citoslovce	CapsLock nebo jiné zvýraznění
Text 1	10	10	10	7	8	8	1	2	10	0	0	0	3	3
Text 2	7	5	0	10	8	10	0	3	10	2	0	0	0	3
Text 3	7	10	0	0	8	7	5	2	10	0	0	0	0	0
Text 4	5	9	0	0	0	5	10	5	2	0	0	0	0	0
Text 5	5	3	0	10	2	3	2	8	5	10	1	8	5	3
Text 6	7	10	5	10	10	5	0	2	8	0	0	0	1	0
Text 7	5	10	0	1	1	0	7	5	8	0	0	0	0	0
Text 8	10	10	5	10	10	3	3	3	10	0	0	0	3	2
Text 9	7	10	0	1	0	0	5	2	8	10	2	0	0	0
Text1	10	10	10	10	0	10	0	5	10	0	0	0	0	0
Text	10	10	0	0	5	0	10	5	3	0	0	0	0	1
Text	10	10	0	0	0	10	10	2	5	8	0	0	3	0
Text	8	9	10	10	0	1	1	1	1	1	0	0	1	3
Text	10	10	0	0	10	0	2	2	8	0	0	0	0	0
Text	10	10	10	10	8	0	3	8	5	0	1	0	0	1
Text	10	10	0	3	2	10	1	2	3	0	0	0	0	0
Text	10	10	0	0	5	0	10	2	5	0	0	0	0	0
Text	7	3	3	10	10	0	1	2	3	0	0	10	0	0
Text	10	10	0	0	10	0	7	1	2	1	0	0	0	0
Text 20	10	9	10	10	10	0	0	4	5	0	0	0	1	3

Kritéria/Texty	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8	Text 9	Text 10	Text 11	Text 12	Text 13	Text 14	Text 15	Text 16	Text 17	Text 18	Text 19	Text 20
Interpunkce	10	7	7	5	5	7	5	10	7	10	10	10	8	10	10	10	10	7	10	10
A/a	10	5	10	9	3	10	10	10	10	10	10	10	9	10	10	10	10	3	10	9
Emotikony	10	0	0	0	0	5	0	5	0	10	0	0	10	0	10	0	0	3	0	10
Počet teček	7	10	0	0	10	10	1	10	1	10	0	0	10	0	10	3	0	10	0	10
Oslovování	8	8	8	0	2	10	1	10	0	0	5	0	0	10	8	2	5	10	10	10
Cizí slova	8	10	7	5	3	5	0	3	0	10	0	10	1	0	0	10	0	0	0	0
Gramatické chyby	1	0	5	10	2	0	7	3	5	0	10	10	1	2	3	1	10	1	7	0
Nespisovnost	2	3	2	5	8	2	5	3	2	5	5	2	1	2	8	2	2	2	1	4
Komplexita věty	10	10	10	2	5	8	8	10	8	10	3	5	1	8	5	3	5	3	2	5
Konfliktnost	0	2	0	0	10	0	0	0	10	0	0	8	1	0	0	0	0	0	1	0
Nadávky	0	0	0	0	1	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0
Rasismus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
Citoslovce	3	0	0	0	5	1	0	3	0	0	0	3	1	0	0	0	0	0	0	1
CapsLock nebo jiné zvýraznění	3	3	0	0	3	0	0	2	0	0	1	0	3	0	1	0	0	0	0	3

Pro lepší přehlednost a zjednodušení práce s tabulkami jsme si Tabulka_1 transponovali do Tabulka_2.

Z výsledků získaných v Tabulka_1 je patrné, že všech pět autorů klade důraz na používání interpunkce. Všichni bez výjimky píší ve svých textech háčky a čárky. Někteří získali méně bodů, protože jim chyběly čárky v souvětích, výjimečně pak tečky za větou. Stejně tak je patrné, že se ve všech textech s výjimkou Text 2, Text 5 a Text 18 nevyskytovaly chyby v používání velkých a malých písmen na začátku věty ani u vlastních jmen. Tato kritéria tedy nejsou pro naši analýzu příliš přínosná.

Dále je z výsledků získaných v Tabulka_1 patrné, že se autoři v používání emotikonů velice liší. Texty číslo 1, 10, 13, 15 a 20 vykazují velké množství emotikonů, což by mohlo poukazovat na to, že byly napsány jedním autorem. Texty číslo 6, 8 a 18 obsahují menší množství emotikonů, což by mohlo znamenat, že se jedná o stejného autora jako texty s hodnocením 10, nebo o dalšího autora. Ostatní texty neobsahovaly žádné emotikony. Jako příklad uvádíme úryvek z Textu 15:

Mučitelé koní ať se smaží v pekle,kde je v předsíni budou ožírat mouchy:-(

Stejně jako v používání emotikonů se autoři textů zásadně liší v používání teček za větami. Autoři Textů číslo 1, 2, 5, 6, 8, 13, 15, 18 a 20 mají ve zvyku psát za některými větami tři tečky, ve vzácnějších případech pak dvě tečky. Jako příklad uvádím úryvek z Textu 18:

Koňové mají krásné oči....mimochodem, taky pštrosové... A prý se na pštrosech dá i jezdit... ???? .... neviděl jsem, slyšel jsem... obhajovat to nebudu...

Jednotliví autoři se výrazně liší v tom, v jaké míře ve svých textech používají oslovování. Autoři komentářů se často odkazují na sebe navzájem a reagují na komentáře ostatních uživatelů. Jako příklad uvádím úryvek z Textu 6:

Dokonale s Vámi souhlasím, pane Ivane. Dokonce i V.Č. si myslí, že dominujícím jazykem nebude čeština. Jenom bych si to přál.

Autoři Textů číslo 1, 2, 3, 6, 8, 14, 15, 18, 19 a 20 používají oslovování velice často. Můžeme pozorovat, že se ve většině případů jedná o texty, ve kterých se objevují dvě a více teček za větou. V případě Textů 1, 6, 8, 15, 18 a 20 můžeme pozorovat, že autoři splňují všechna tři kritéria: obsahují emotikony, věty ukončené více než jednou tečkou a zároveň se v nich objevuje oslovování ostatních uživatelů. Jako příklad uvádíme úryvek z Textu 8:

Omluva, pane Čechu. Holt ty germanismy musím nějak okecat,když babička byla germánka.....;-)

Na základě interpretace dat lze předpokládat, že by se mohlo jednat o jednoho nebo dva autory. Kritéria emotikony, množství teček za větou a oslovování tedy pro naši analýzu hodnotíme jako přínosné.

Jako další hodnotící kritérium jsme si zvolili používání cizích slov. V Tabulce_1 můžeme vidět, že polovina autorů cizí slova používá a polovina nepoužívá. Jako příklad uvádím úryvek z Textu 2:

A konec konců, četl jsem úvahu, že k porážce Luftwaffe přispělo dosti paradoxně i použití raket V2, když jejich výroba odčerpala zdroje na stavbu bombardovacích letadel, které Británii ničila DALEKO levněji…

Toto kritérium tedy není pro naši analýzu příliš přínosné.

Mezi nejužitečnější kritéria, která mohou pomoci s identifikací autora, by mohl být zařazen výskyt gramatických chyb v textu. Často se jedná o chyby ve shodě podmětu s přísudkem nebo v rozlišování měkkého a tvrdého i/y. Texty číslo 4, 7, 11, 12, 17, 19 obsahují výrazně více gramatických chyb než texty zbývající. Jako příklad uvádíme úryvek z Textu 17:

Jenže přesně stejně si to myslí i ti, co dělaly tyto hrozné činy. Platí to heslo: „kdo jsi bez chyby hoď kamenem“. Každý by především měl spytovat své svědomí.

Na základě interpretace dat lze předpokládat, že by se mohlo jednat o jednoho nebo dva autory.

V Tabulka_1 můžeme vidět, že používání nespisovného jazyka není kritériem, které by od sebe výrazně odlišovalo jednotlivé zkoumané texty či autory a na jeho základě tedy není možné odhadovat, které texty patří k sobě. Téměř ve všech případech získaly texty méně než 5 bodů, z čehož vyplývá, že autoři používají spíše spisovný jazyk. Výjimkou jsou Text 5 a Text 15, které shodně získaly 8 bodů. Jako příklad uvádíme úryvek z Textu 15:

A lidi raději žijousami,než aby si vzali domů věrnýho kamaráda. Je to dané výchovou,nemůžebejtkaždejzvířatomil.

V Tabulka_1 můžeme pozorovat, že kritéria konfliktnosti, rasismu a projevů nenávisti a používání nadávek by mohla být využita k rozlišení jednotlivých autorů. Jako konfliktní působí Texty 2, 5, 9, 12, 13 a 19, rasovou nebo náboženskou nenávist vykazovaly Texty 5 a 18 a nadávky se objevily v Textech 5, 9 a 15. Jako příklad uvádíme ukázku z Textu 5, který vykazuje všechny tři výše zmíněné znaky:

Nehledě na bolest a jiné trápení (z lenosti a hlupáctví vynázce tohoto postupu) raněného účastníka této pyramidální pitomosti. Inu, v telecím mládí vynalézají mlamojovéneuvěřitelné hlouposti. Kolik let bylo těm Patům a Matům, kteří zmrzačili svým postupem spolupracovníka?

Na základě interpretace dat lze tedy předpokládat, že by Texty 2, 5, 9, 12, 13, 15, 18 a 19 mohly patřit dvěma autorům.

Kritérium komplexita věty nám rozděluje autory na dvě poloviny. Autoři Textů 1, 2, 3, 6, 7, 8, 9, 10, 11 a 15 užívají bohatou syntax. Věty jsou delší

a vyskytuje se v nich velké množství vložených a vedlejších vět. Jako příklad uvádíme úryvek z Textu 1, který za toto kritérium získal 10 bodů:

> *Zatímco před necelými pár sto lety by babka zaříkačka zaříkávala a oko čadila léčivým bylím,pilo by se bílý víno s naloženejmapetrklíčema a koupalo by se ve Světlíku,nechalo by se oko,aby pracovalo samo podle svého,dnes to je podle diktátu nejposlednější vědy a ještě s berlou Mrazilkou!*

Toto kritérium není příliš užitečné k oddělení jednotlivých autorů od sebe.

Poslední dvě kritéria, jež jsme si zvolili, sledovala, v jaké míře jednotliví autoři používají citoslovce a jak často ve svém projevu využívají CapsLock nebo nějaký jiný způsob zvýraznění textu. Citoslovce obsahují Texty 1, 5, 6, 8, 12, 13 a 20. Zvýraznění slov obsahují Texty 1, 2, 5, 8, 11, 13, 15 a 20. V případě Textů 1, 5, 8, a 20 můžeme pozorovat, že autoři splňují obě dvě kritéria zároveň, z čehož by mohlo vyplývat, že se jedná o stejného autora.

V této části jsme se pokusili navrhnout několik hypotéz ohledně toho, které texty by na základě výsledků v Tabulka_1 a Tabulka_2 mohly patřit ke stejným autorům. Hypotézy jsou tedy následující:

Texty číslo 6, 8 a 18 obsahují menší množství emotikonů, což by mohlo znamenat, že se jedná o stejného autora jako texty s hodnocením 10, nebo o dalšího autora.

V případě Textů 1, 6, 8, 15, 18 a 20 můžeme pozorovat, že autoři splňují všechna tři kritéria, tedy jejich texty obsahují emotikony, věty ukončené více než jednou tečkou a zároveň se v nich objevuje oslovování ostatních uživatelů. Na základě interpretace dat lze předpokládat, že by se mohlo jednat o jednoho autora, nebo o dva autory.

Texty číslo 4, 7, 11, 12, 17, 19 obsahují výrazně více gramatických chyb než texty zbývající. Na základě interpretace dat lze předpokládat, že by se mohlo jednat o jednoho autora, nebo o dva autory.

Na základě kritérií konfliktnosti, rasismu a projevů nenávisti a používání nadávek lze předpokládat, že by texty 2, 5, 9, 12, 13, 15, 18 a 19 mohly patřit dvěma autorům.

V případě Textů 1, 5, 8, a 20 můžeme pozorovat, že autoři splňují kritérium používání funkce CapsLock a zároveň velkého množství citoslovcí, z čehož by mohlo vyplývat, že se jedná o stejného autora.

Dále jsme vytvořili Tabulka_souvislosti_3, která měla zjednodušit orientaci v hodnocení textů, nebyla však příliš užitečná, jelikož je velice obsáhlá a pro hodnotícího nepřehledná. Existují však metody, které dokážou převést tabulku na mapu podobnosti jednotlivých autorů a umožňují nám porovnávat texty mezi sebou kvantitativním a objektivním způsobem. Tyto metody se označují jako vícerozměrné škálování a hierarchické shlukování. Tyto metody budou představeny v následující kapitole.

	9	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**
11,12,14,15,16,17,19,20		13	2,3,6,18		4,5,7,					9
,8,9,10,11,12,14,15,16,	4,13,20				2		5,18			
,15,20					6,8		18			2,3,4,5,7,9,11,12,14,1
,10,13,15,18,20			1				16		7,9	3,4,11,12,14,17,19
18,19,20		1,2,3,15			11,17			5,16	7	4,9,10,12,13
,16		1	3		4,6		5,8		13	7,9,11,14,15,17,18,19
,17			7,19		3,9		8,15	5,14	1,13,16,18	2,6,10,20
		5,15			4,7,10,11	20	2,8	1,3,6,9,12,14,16,17,18	13,19	
,10		6,7,9,14			5,12,15,17,20		11,16,18	4,19	13	
		12						2	13,19	1,3,4,6,7,8,10,11,14,1
								9	5,15	1,2,3,4,6,7,8,10,11,12,20
										1,2,3,4,5,6,7,8,9,10,1,19,20
					5		1,8,12		6,13,20	2,3,4,7,9,10,11,14,15
							1,2,5,13,20	8	11,15	3,4,6,7,9,10,12,14,16

osti_3

6.2. Sumarizace

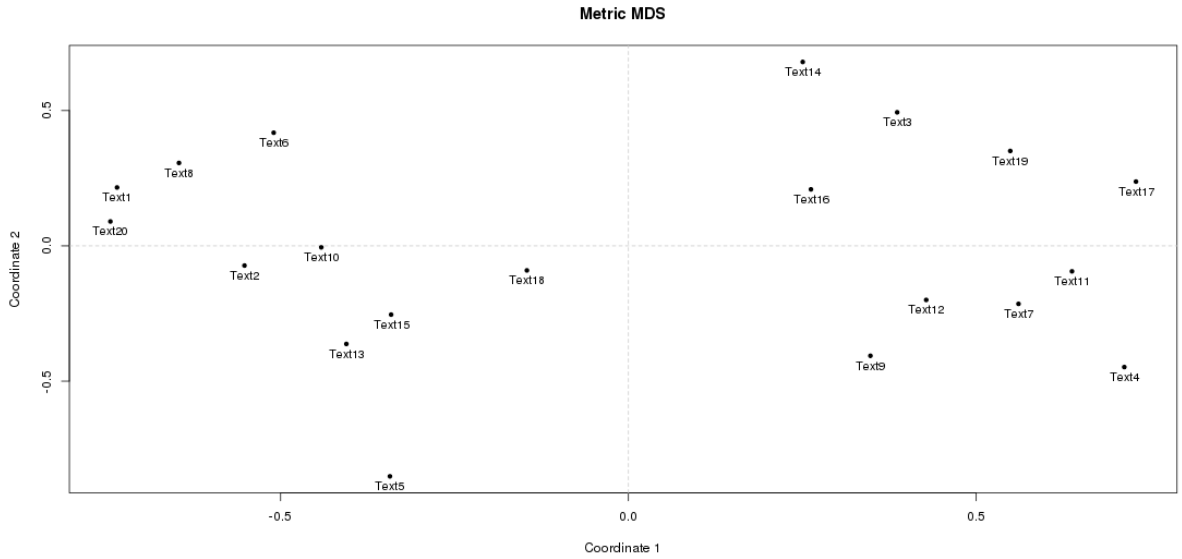
Jako hodnotící kritéria jsme si zvolili čtrnáct znaků jazykového vyjadřování, které bývají pro jednotlivé autory typické. Těmito kritérii jsou interpunkce, používání velkých a malých písmen, citoslovcí, emotikonů, výskyt dvou a více teček za větou, oslovování, výskyt cizích slov, výskyt gramatických chyb, nespisovného jazyka a nadávek. Dále jde o komplexitu vět, konfliktnost, rasismus a projevy nenávisti vůči minoritním skupinám, používání CapsLocku nebo jiného zvýraznění slov. Jednotlivá kritéria jsme na základě kvalifikovaného odhadu ohodnotili na škále od 0 do 10, 10 znamená nejvíc a nula úplně chybí.

6.3. Vícerozměrné škálování (multidimensional scaling)

Díky vícerozměrnému škálování jsme schopni porovnávat texty objektivním způsobem a jsme schopni je vizualizovat. Data se na základě kvantifikovaných vlastností shlukují do skupin, což nám dává možnost je na základě vlastností přiřadit k sobě. [20]

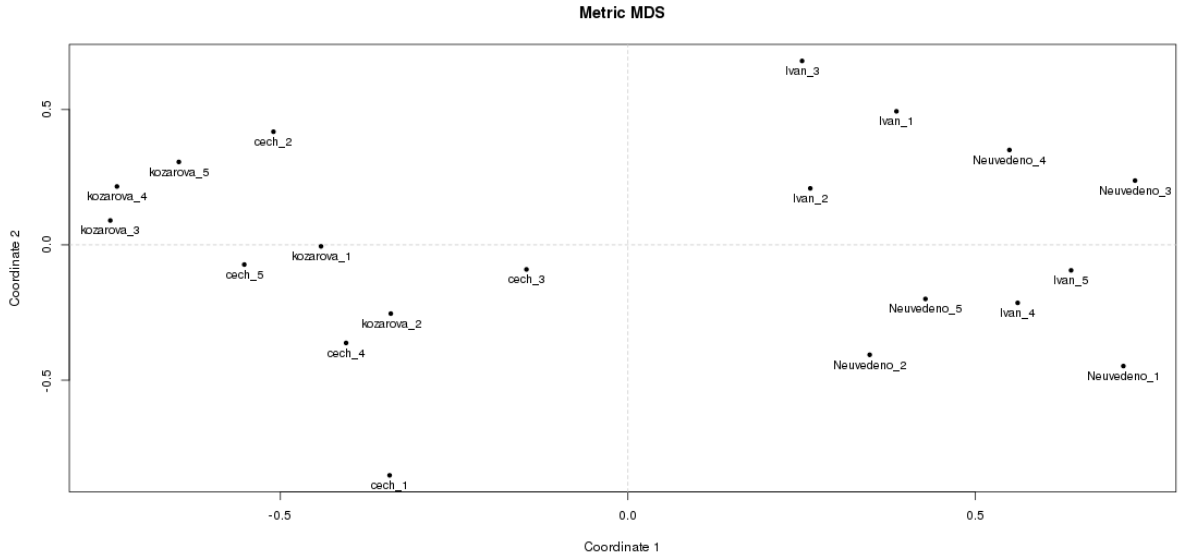
Při analýze používáme euklidovskou vzdálenost a kosinovou podobnost. Rozdíl mezi těmito dvěma metodami spočívá v tom, že euklidovská vzdálenost bere v potaz celkové množství, kterým se dvě kritéria odlišují, zatímco kosinová podobnost (v našem případě spíše nepodobnost) rozdíly v množství nebere v úvahu a její výsledek odpovídá tomu, jak jednotlivá kritéria korespondují společně, tedy zachovávají shodné poměry. Pro více detailů viz kniha *Od kvantitativní lingvistiky k neuronovým sítím: Přístupy k analýze textu*.

[20] FALTÝNEK, Dan, Dalibor PAVLAS, Ondřej VRABEL a Vladimír MATLACH. *Od kvantitativní lingvistiky k neuronovým sítím: Přístupy k analýze textu*. Olomouc: Univerzita Palackého, 2015, kniha je v editaci.



Obrázek 1: Graf výsledků vícerozměrného škálování za použití kosinové podobnosti.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. Můžeme si všimnout dvou výrazně oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části. Texty byly rozděleny přesně na poloviny, v pravé části obrázku se nachází shluk deseti textů a v levé části také shluk deseti textů, tyto shluky však není možné dále oddělit.

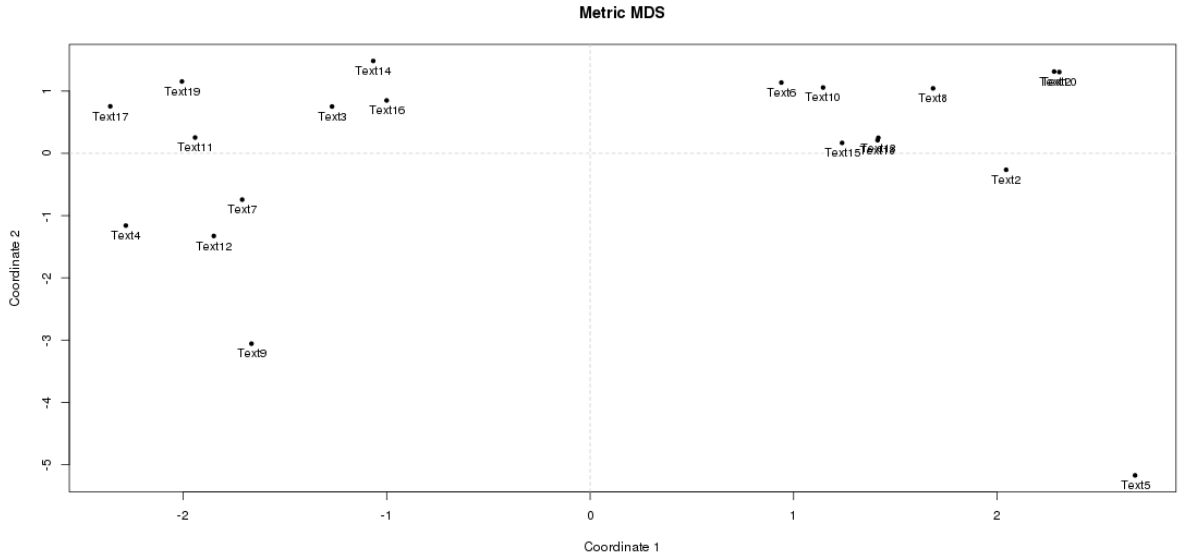


Obrázek 2: Graf výsledků vícerozměrného škálování za použití kosinové podobnosti.

$GOF[1] 0.4278800

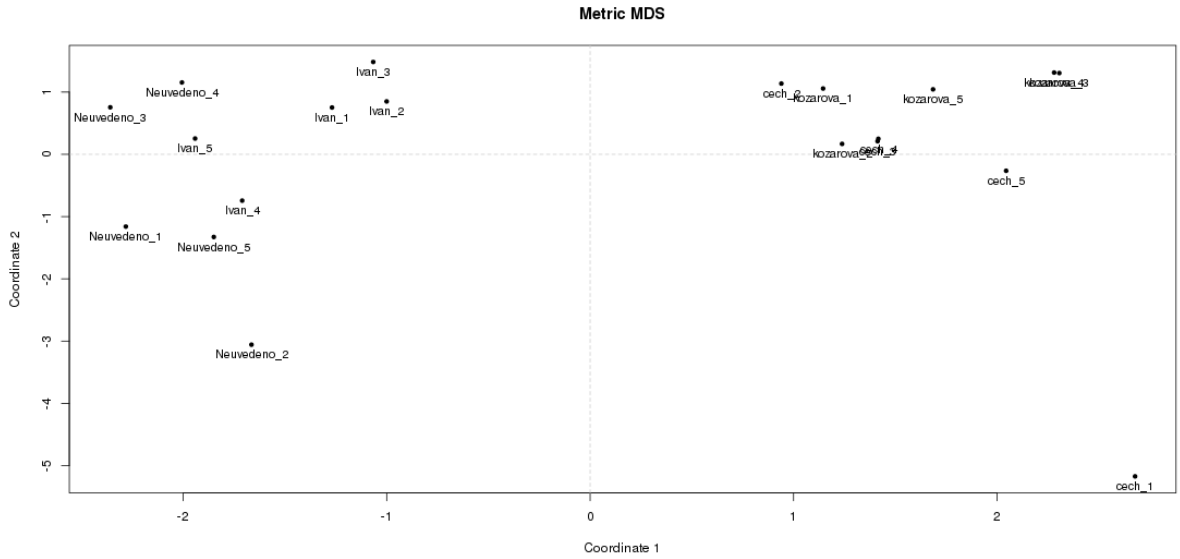
Můžeme si všimnout dvou oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části. Lze pozorovat, že v pravé části se nachází texty od autorů Ivan a Neuvedeno, v levé části se pak nachází texty od autorů Kozárová a Čech. Jednotlivé autory nelze oddělit lineárně, ale je možné pozorovat, že se texty jednotlivých autorů shlukly blízko sebe.

V případě této analýzy MDS zrekonstruovalo přibližně 43 % variance vypočítaných vzdáleností, což znamená, že některé informace o podobnosti či vzdálenosti zadaných textů nejsou vidět. I přesto můžeme v grafu vidět shluky, které odpovídají jednotlivým autorům.



Obrázek 3: Graf výsledků vícerozměrného škálování za použití euklidovské vzdálenosti.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. Můžeme si všimnout dvou výrazně oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části obrázku. Opět se texty rozdělily na shluky po deseti, ve kterých není možné přesně oddělit jednotlivé autory od sebe. Text číslo 5 byl vyhodnocen jako nejvíce odlišný, a proto leží v pravém dolním rohu nejvíce vzdálený od všech ostatních textů.



Obrázek 4: Graf výsledků vícerozměrného škálování za použití euklidovské vzdálenosti

$GOF[1] 0.4369482

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. Můžeme si všimnout dvou jasně oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části obrázku. Opět není možné autory lineárně oddělit, ale můžeme pozorovat, že se shlukli do skupin. Výjimkou je Čech_1, který se nachází v pravém dolním rohu a tvoří tak samostatnou skupinu.

V případě této analýzy MDS zrekonstruovalo přibližně 43 % variance vypočítaných vzdáleností, což znamená, že některé informace o podobnosti či vzdálenosti zadaných textů nejsou vidět. I přesto můžeme v grafu vidět shluky, které odpovídají jednotlivým autorům.

6.3.1. Závěr

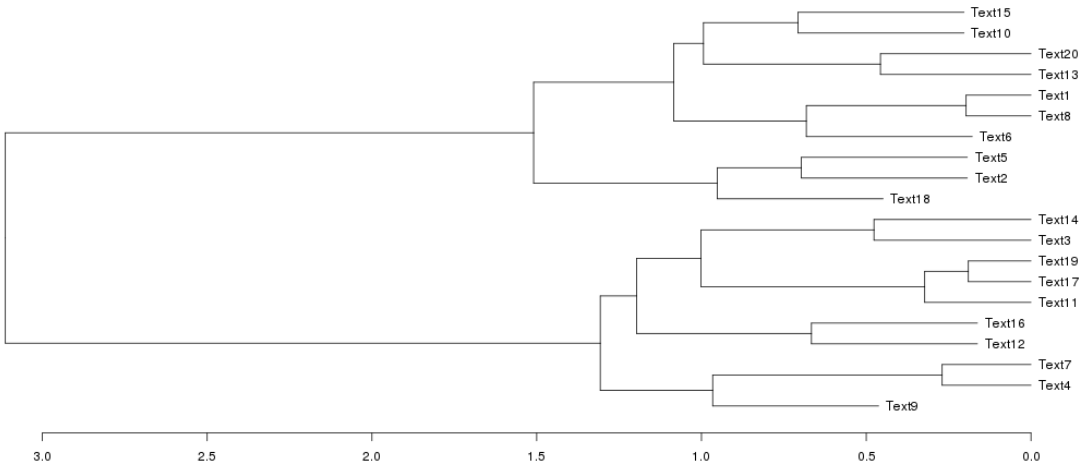
Interpretací jednotlivých grafů vícerozměrného škálování jsme zjistili, že lze oddělit dvě skupiny výrazně odlišných textů – to může svědčit o dvou autorech, nebo o dvou skupinách autorů.

6.3.2. Sumarizace

V této kapitole bylo čtenáři představeno vícerozměrné škálování. Tato metoda nám umožňuje vytvořit graf, na kterém můžeme pozorovat jednotlivé shluky textů a na jejich základě se pokusit odhadnout, které texty patří kterému autorovi. Dále byly představeny a vysvětleny pojmy euklidovská vzdálenost a kosinová podobnost.

6.4. Hierarchické shlukování

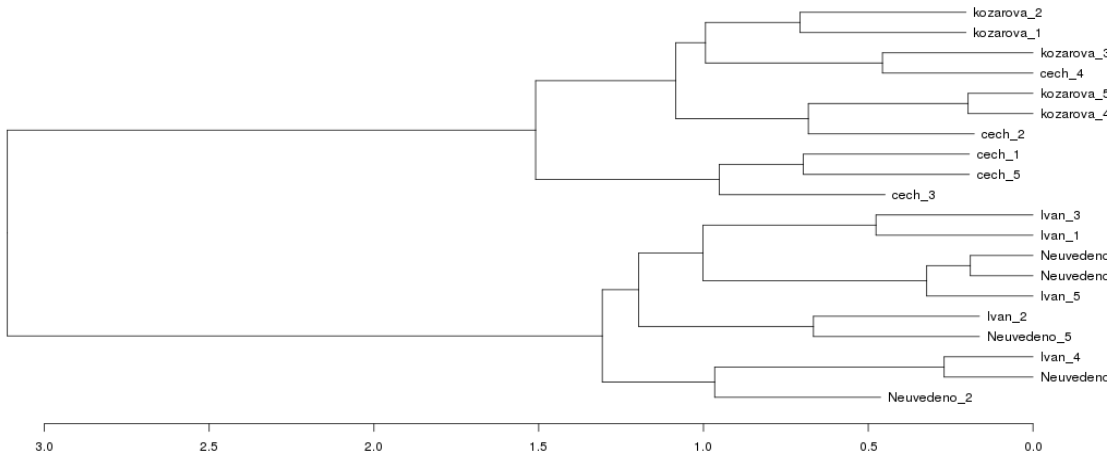
„O něco komplexnější a zajímavější náhled na data nám nabízí metoda tzv. hierarchického shlukování a jeho grafická interpretace formou tzv. dendrogramu (stromového diagramu). Hierarchické shlukování k sobě na základě matice vzdáleností přiřazuje vždy dva nejpodobnější (nejbližší) objekty (texty), které jsou v dendrogramu zobrazeny na stejné spojnici (jsou jedinými dvěma listy stejné větve).“[21]



Obrázek 5: Graf výsledků hierarchického shlukování za použití kosinové podobnosti.

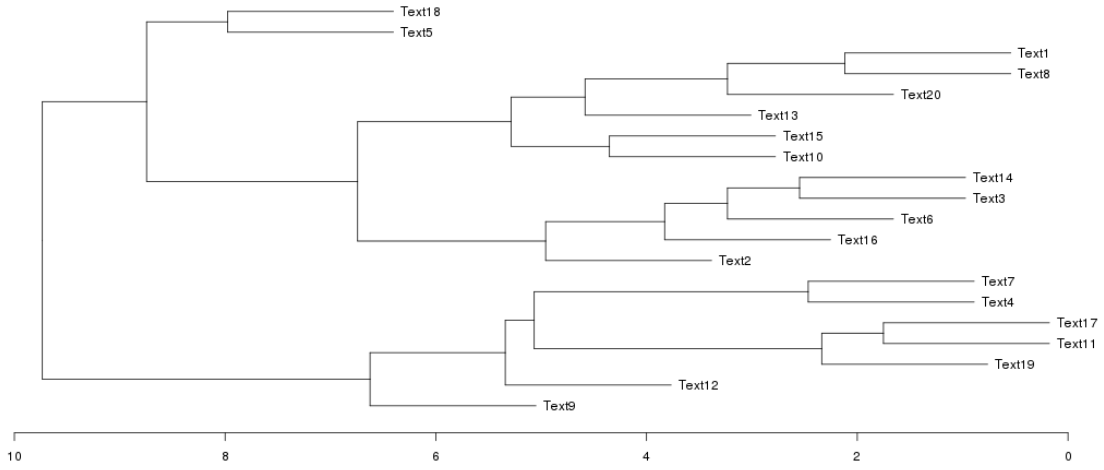
Hierarchické shlukování vytvořilo náhled na podobnost jednotlivých textů. Můžeme si všimnout čtyř jasně oddělených větví. Autory se nám nepodařilo přesně oddělit, protože v horní polovině grafu obsahuje jedna větev sedm textů a druhá pouze tři, ve spodní části grafu taktéž obsahuje jedna větev sedm textů a druhá pouze tři.

[21] FALTÝNEK, Dan, Dalibor PAVLAS, Ondřej VRABEL a Vladimír MATLACH. *Od kvantitativní lingvistiky k neuronovým sítím: Přístupy k analýze textu*. Univerzita Palackého, 2015 kniha je v editaci.



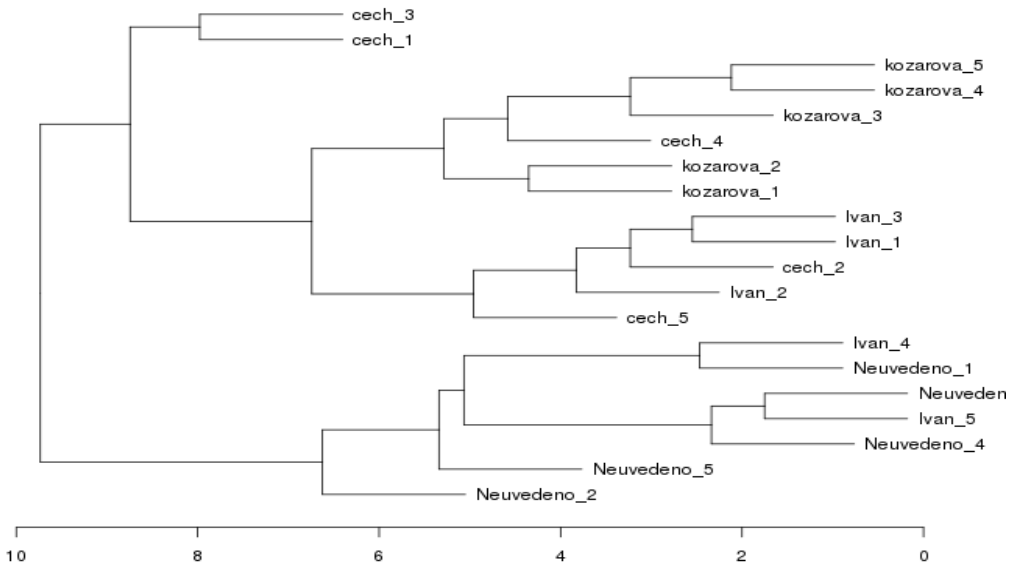
Obrázek 6: Graf výsledků hierarchického shlukování za použití kosinové podobnosti.

Hierarchické shlukování vytvořilo náhled na podobnost jednotlivých textů. Můžeme vidět čtyři jasně oddělené shluky textů. Autory se nám nepodařilo přesně oddělit, ale můžeme pozorovat, že v horní polovině grafu se vyskytují Kozárová a Čech a až na jeden případ se na spojnicích setkal stejný autor. Ve spodní polovině grafu se shlukli autoři Ivan a Neuvedeno, jejichž oddělení není tak jednoznačné jako u Kozárové a Čecha.



Obrázek 7: Graf výsledků hierarchického shlukování za použití euklidovské vzdálenosti.

Hierarchické shlukování vytvořilo náhled na podobnost jednotlivých anonymních textů. Můžeme pozorovat, že rozdělení na jednotlivé větve je komplikovanější než u hierarchického shlukování s použitím kosinové vzdálenosti. Jednotlivé struktury jsou složitější a není možné je snadno oddělit na menší celky. Můžeme odhadovat, že Text 5 a Text 18 patří stejnému autorovi, protože jsou jasně odděleny od ostatních. Eukleidovská vzdálenost nám v tomto případě neposkytuje příliš jasné výsledky a na základě klastrování není možné identifikovat skupiny textů příslušející jednomu autorovi (nebo jedné skupině autorů), jak tomu bylo v případě kosinové vzdálenosti.



Obrázek 8: Graf výsledků hierarchického shlukování za použití euklidovské vzdálenosti.

Hierarchické shlukování vytvořilo náhled na podobnost jednotlivých pojmenovaných textů. Potvrdilo se, že Text 5 a Text 18 patří stejnému autorovi, ostatní struktury jsou velice komplikované. Je možné pozorovat, že autoři Čech a Kozárová se nachází v těsné blízkosti v horní části grafu, v prostřední části grafu se shlukli Ivan a Čech a ve spodní části grafu se nachází Ivan a Neuvedeno. Oddělení Kozárové dopadlo úspěšně, protože veškeré její texty tvoří jeden shluk.

6.4.1. Závěr

Hierarchické shlukování a způsob výpočtu vzdáleností identifikovaly celkem úspěšně autory Kozárová a Neuvedeno, protože jejich texty tvoří dva samostatné shluky jen s nízkou kontaminací, ostatní autoři jsou buď promíchaní, nebo vzdálení od sebe sama.

Při odhalení autorů se nám potvrdilo, že eukleidovská vzdálenost nám v tomto případě neposkytuje příliš jasné výsledky a na základě klastrování není možné identifikovat skupiny textů příslušející jednomu autorovi (nebo jedné skupině autorů), jak tomu bylo v případě kosinové vzdálenosti.

6.4.2. Sumarizace

V této kapitole jsme popsali metodu hierarchického shlukování. Výsledkem hierarchického shlukování je tzv. dendrogram, stromový diagram. Na jednotlivých spojnicích jsou vždy přiřazeny dva nejpodobnější texty.

6.5. Závěr testování hypotéz

V kapitole 6.2. jsme se pokusili navrhnout několik hypotéz, abychom se pokusili zjistit, které texty by na základě výsledků v Tabulka_1 a Tabulka_2 mohly patřit ke stejným autorům. Nyní na základě výsledků hierarchického shlukování a vícerozměrného škálování můžeme vyhodnotit, jestli je možné hypotézy potvrdit nebo vyvrátit.

Texty číslo 6, 8 a 18 obsahují menší množství emotikonů, což by mohlo znamenat, že se jedná o stejného autora jako u textů s hodnocením 10, nebo o dalšího autora. Texty číslo 6, 13 a 18 patří Čechovi, Texty číslo 1, 8, 10, 15 a 20 Kozárové. Hypotézu se nám tedy podařilo potvrdit. Můžeme tedy vidět, že emotikony používají v komunikaci 2 ze 4 autorů, Kozárová dokonce ve všech svých pěti textech.

V případě Textů číslo 1, 6, 8, 15, 18 a 20 můžeme pozorovat, že autoři splňují všechna tři kritéria, obsahují tedy emotikony, věty ukončené více než jednou tečkou a zároveň se v nich objevuje oslovování ostatních uživatelů. Na základě interpretace dat lze předpokládat, že by se mohlo jednat o jednoho nebo dva autory. Texty číslo 6 a 18 patří Čechovi, Texty číslo 1, 8, 15 a 20 Kozárové. Hypotézu se nám tedy podařilo potvrdit.

Texty číslo 4, 7, 11, 12, 17, 19 obsahují výrazně více gramatických chyb než texty zbývající. Na základě interpretace dat lze předpokládat, že by se mohlo jednat o jednoho nebo dva autory. Texty číslo 4, 7, 12, 17 a 19 patří autorovi Neuvedeno, můžeme tedy pozorovat, že autor se dopouští gramatických chyb v každém ze svých textů. Text číslo 11 patří autorovi Ivan. Hypotézu se nám tedy podařilo potvrdit.

Na základě kritérií konfliktnosti, rasismu a projevů nenávisti a používání nadávek lze předpokládat, že by Texty číslo 2, 5, 9, 12, 13, 15, 18 a 19 mohly patřit dvěma autorům. Texty číslo 2, 5, 13 a 18 patří Čechovi. Vidíme tedy, že autor Čech je agresivní ve čtyřech ze svých pěti textů. Texty číslo 9, 12 a 19 patří autorovi Neuvedeno, který je tudíž agresivní ve třech z pěti textů. Hypotézu se nám ovšem nepodařilo potvrdit, jelikož Text 15, který také získal vysoké bodové ohodnocení, patří autorce Kozárové.

V případě Textů číslo 1, 5, 8, a 20 můžeme pozorovat, že autoři splňují kritérium používání CapsLock a zároveň velkého množství citoslovcí, z čehož

by mohlo vyplývat, že se jedná o stejného autora. Tuto hypotézu se nám potvrdit nepodařilo, protože Texty 1, 8 a 20 patří Kozárové a Text 5 Čechovi.

7. Množina slov (bag-of-words model)

V této kapitoly se zaměříme na metodu bag-of-words, vysvětlíme její fungování a provedeme analýzu anonymních textů za pomoci hierarchického shlukování a vícerozměrného škálování. Interpretací jednotlivých grafů se pokusíme odhadnout, které texty by mohly patřit k sobě. Analýzu provedeme znovu pro neanonymní texty a vyhodnotíme, nakolik byl náš odhad správný a zda jsme dosáhli lepších výsledků při použití hierarchického shlukování nebo vícerozměrného škálování.

Při použití bag-of-words modelu není důležité pořadí slov v dokumentu, ale frekvence jejich výskytu.[22] Nemusejí se používat pouze slova samotná, ale mohou se používat i n-tice slov (n-gramy). N-gramy jsou po sobě jdoucí sekvence n prvků.

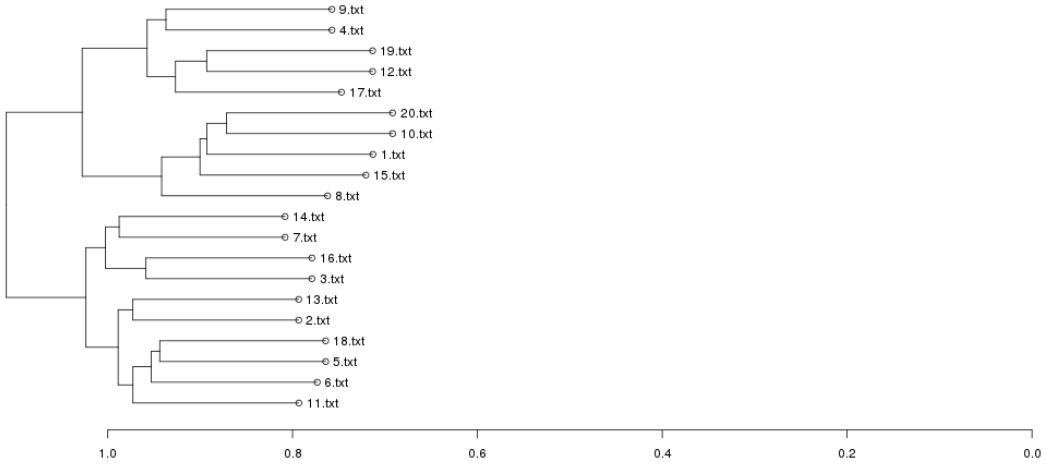
V následující analýze byly použity 2-tice slov, takzvané bigramy. Tento model rozdělí text na dvojice po sobě jdoucích slov. Takto vzniklé řetězce bigramů textů pak vzájemně srovnáváme podle toho, zda se na dané pozici vyskytuje stejný bigram. Texty s největším množstvím shod bigramů na dané pozici jsou pak vyhodnoceny jako nejpodobnější. Jako příklad uvádíme větu z dokumentu Čech_3:

Ti byli zachráněni sovětským ledoborcem.

Bigramy budou vypadat následovně: ti byli, byli zachráněni, zachráněni sovětským, sovětským ledoborcem.

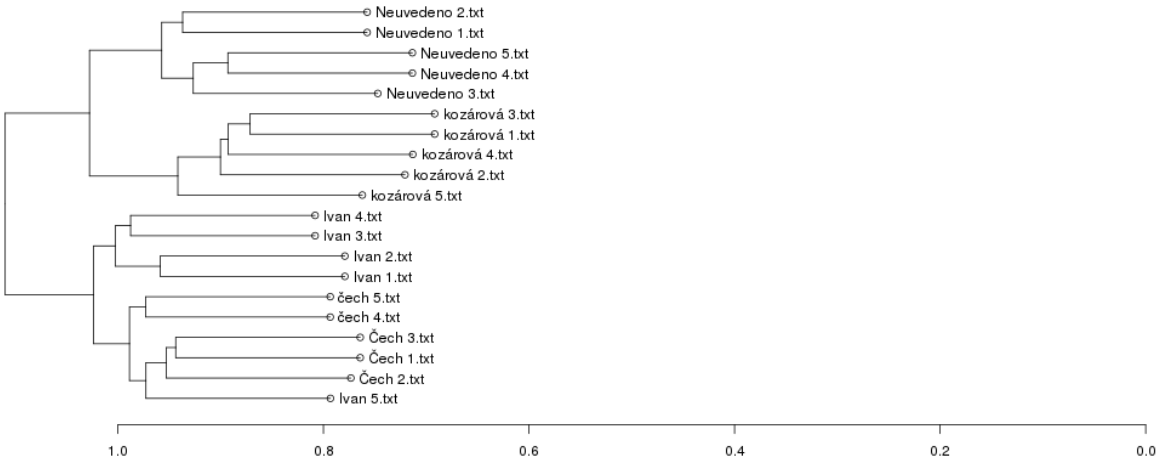
V následující části provedeme analýzu anonymních i neanonymních textů.

[22]Manning, C. D.; Raghavan, P.; Schütze, H.: Introduction to informationretrieval. Cambridge University Press, 2008, ISBN 0521865719.



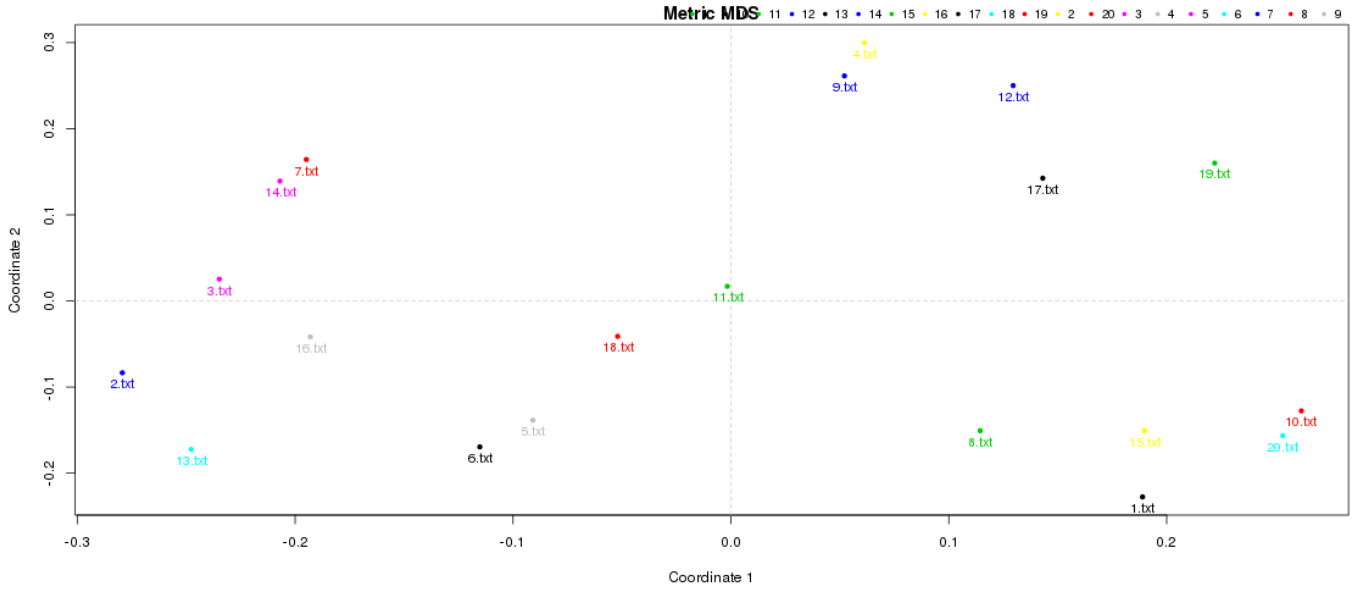
Obrázek 9: Graf výsledků hierarchického shlukování pro 2-gramy slov, kosinová vzdálenost.

Hierarchické shlukování vytvořilo náhled, ve kterém můžeme pozorovat rozdělení na čtyři hlavní větve. Uvnitř těchto čtyř shluků se objevují už jen blízké shluky tvořené jednotlivými texty. V horní polovině grafu se texty shlukly po pěticích, ve spodní polovině grafu se shlukly po čtyřech a šesti textech. Z tohoto důvodu můžeme odhadovat, že minimálně jeden text bude zařazen špatně.



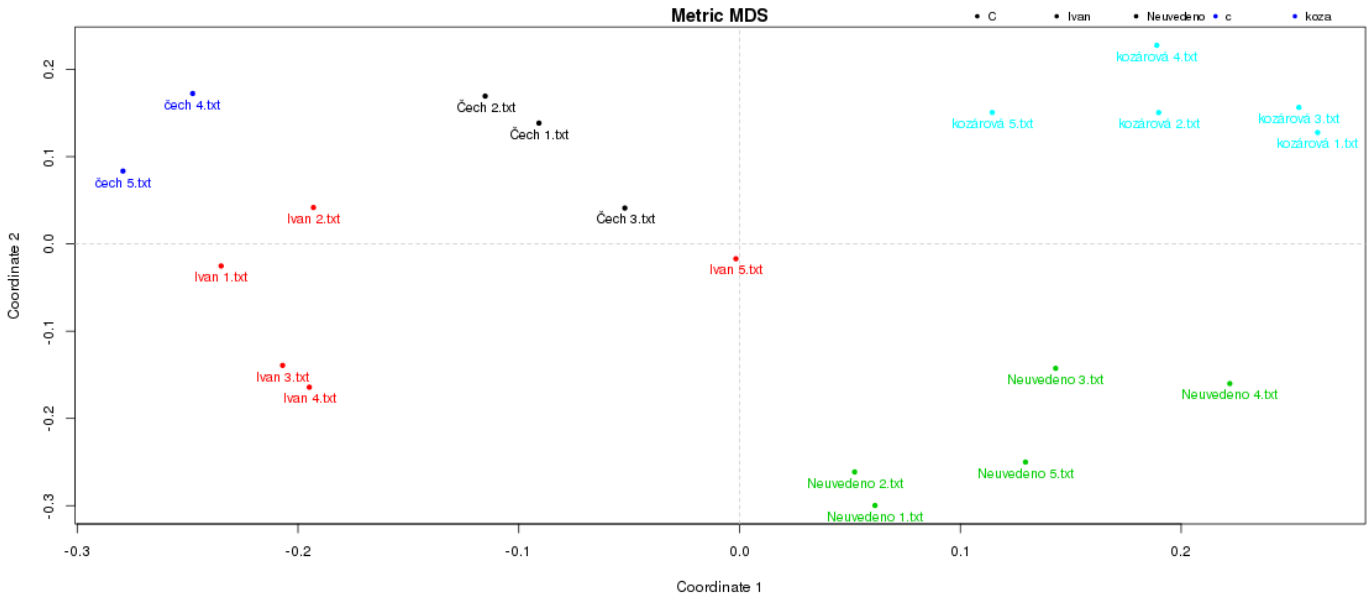
Obrázek 10: Graf výsledků hierarchického shlukování pro 2-gramy slov, kosinová vzdálenost.

Hierarchické shlukování vytvořilo náhled, ve kterém můžeme pozorovat rozdělení na čtyři hlavní větve. Uvnitř těchto čtyř shluků se objevují už jen blízké shluky tvořené jednotlivými texty. V horní části se naprosto jednoznačně shlukly veškeré texty od autora Neuvedeno a pod ním se shlukly veškeré texty autora Kozárová. Tito autoři jsou od sebe jasně odděleni. Ve spodní polovině grafu se také celkem jednoznačně shlukly texty od autorů Ivan a Čech. Potvrdil se náš odhad, že minimálně jeden text bude zařazen špatně, protože jediný text, který se nepřiřadil do správné skupiny, je text Ivan_5, který se zařadil mezi texty autora Čecha.



Obrázek 11: Graf výsledků vícerozměrného škálování pro 2-gramy slov, kosinová vzdálenost.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. Můžeme si všimnout dvou oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části. Můžeme pozorovat, že v pravé části se nachází dvě pětice textů, které se shlukly v opačných rozích grafu, lze tedy odhadovat, že se jedná o dva autory, kteří vykazují nejrozdílnější vlastnosti. V levé části se pak nachází texty, které od sebe nelze jasně oddělit.



Obrázek 12: Graf výsledků vícerozměrného škálování pro 2-gramy slov, kosinová vzdálenost.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. Můžeme si všimnout dvou oddělených shluků textů, jeden v pravé části obrázku a druhý v levé části. Lze pozorovat, že v pravé části se nachází texty od autorů Kozárová a Neuvedeno, autoři Kozárová a Neuvedeno jsou od sebe jasně oddělitelní, texty se shlukly v opačných rozích grafu. V levé části se pak nachází texty od autorů Ivan a Čech. Můžeme pozorovat, že tyto dvě skupiny se nachází velice blízko ke středu, k sobě navzájem a texty jednotlivých autorů jsou promíchány.

7.1. Závěr

Analýza pomocí bag-of-words modelu se ukázala jako velice úspěšná. Obě použité metody se prokázaly jako dobře fungující, hierarchické shlukování však vykazuje o něco lepší výsledky, protože v grafu od sebe můžeme jasně oddělit všechny čtyři skupiny. Vyskytl se v něm pouze jeden špatně zařazený text.

8. Indexy

Jako třetí metodu jsme zvolili hodnocení podle vybraných kvantitativně lingvistických indexů. Jednotlivé indexy budou stručně popsány. Veškeré poznatky v této kapitole vychází z knihy *Metody analýzy (nejen) básnických textů*.[23]

Jednou z typických vlastností každého textu je jeho slovní bohatství. Pro náš výzkum je velice důležité eliminovat vliv délky textu na hodnotu indexu vyjadřujícího slovní bohatství, protože s rostoucí délkou textu roste i velikost slovníku. Proto bude, jak už bylo zmíněno v kapitole Korpus textů, analyzováno celkem dvacet textů od čtyř různých autorů, vždy pět textů o každého autora v rozsahu kolem 5 tisíc znaků. V následující části představíme některé metody, které se využívají k měření slovního bohatství textu.

Entropie

Definicí entropie je míra neurčitosti systému. V případě analýzy frekvenční distribuce slov v textu entropie vyjadřuje hodnotu míry diverzity – čím je hodnota entropie větší, tím diverzifikovanější je slovník, z čehož vyplývá, že čím je vyšší hodnota entropie, tím větší je bohatství slovníku.

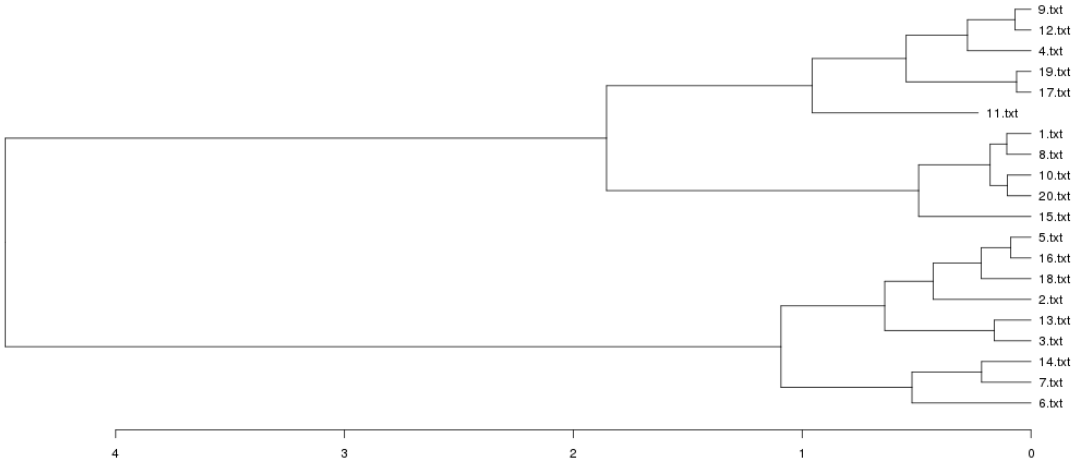
Index opakování slov

Dalším způsobem měření slovního bohatství textu je výpočet indexu opakování slov RR (repeat rate), který vyjadřuje míru koncentrovanosti textu vzhledem k použitému lexiku. Čím je hodnota RR vyšší, tím je menší slovní bohatství textu.

Poměr typů a tokenů (TTR) s

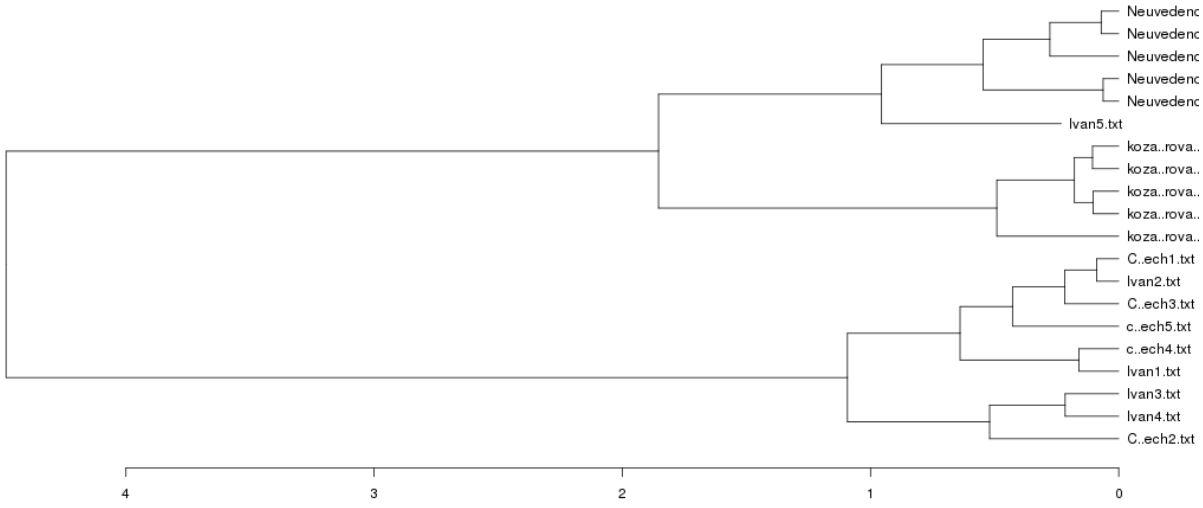
TTR (*type-token ratio)* vyjadřuje poměr počtu různých slov, tzv. typů, k počtu všech slov vyskytujících se v textu, tzv. tokenů. Vyšší *TTR* značí méně opakujících se výrazů, a tedy větší bohatství slovníku.

[23]ČECH, Radek, Ioan-Iovitz POPESCU a Gabriel ALTMANN. *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc: Univerzita Palackého v Olomouci, 2014. Qfwfq. ISBN 978-80-244-4044-6.



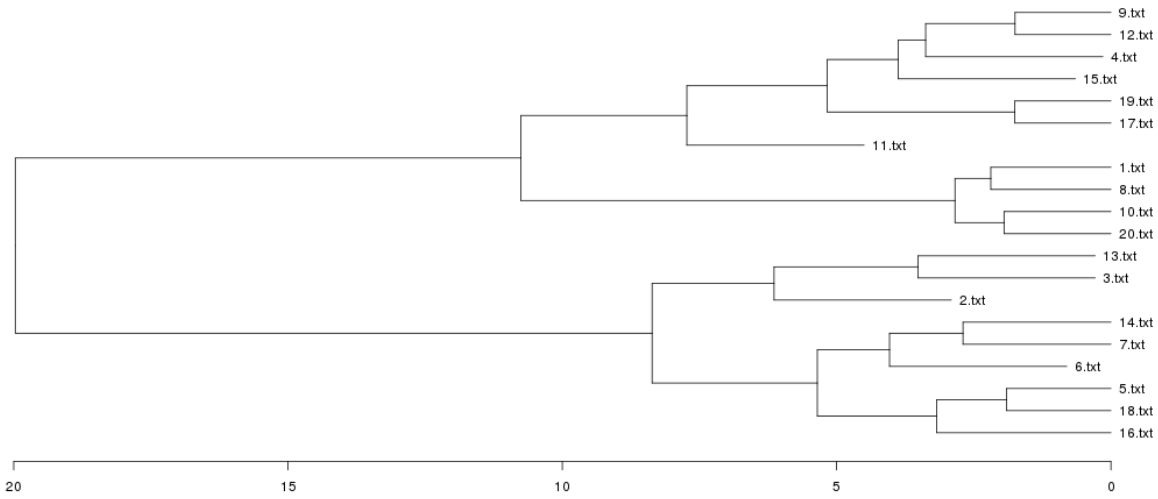
Obrázek 13: Graf výsledků hierarchického shlukování za použití indexů, kosinová podobnost.

Hierarchické shlukování vytvořilo náhled, ve kterém můžeme pozorovat rozdělení na čtyři hlavní větve. Jednotlivé struktury jsou složité. Můžeme odhadovat, že se v horní polovině grafu vyskytnou dva autoři a ve spodní polovině grafu také dva. Dále můžeme odhadovat, že minimálně jeden z textů bude zařazen špatně, protože v horní polovině grafu se shluklo textů jedenáct a ve spodní polovině jen devět.



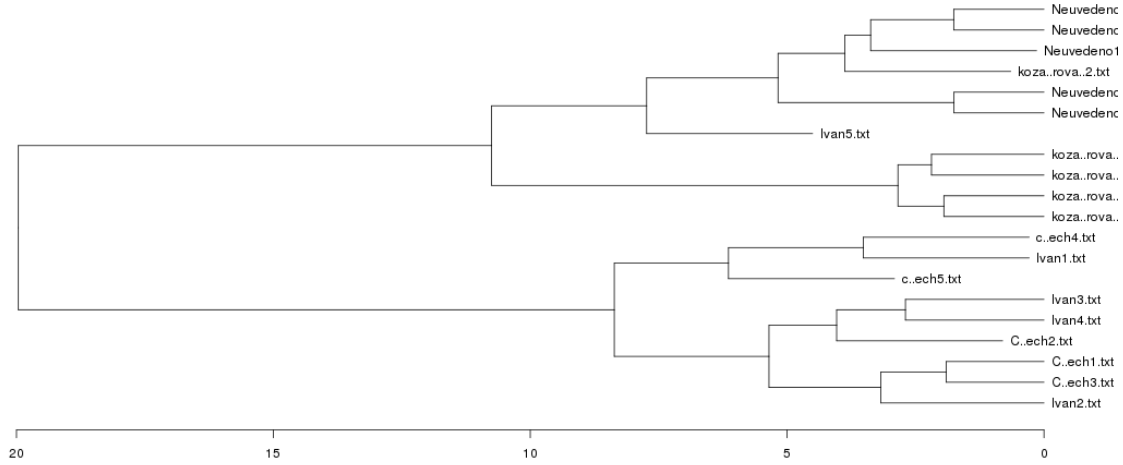
Obrázek 14: Graf výsledků hierarchického shlukování za použití indexů, kosinová podobnost.

Hierarchické shlukování vytvořilo náhled, ve kterém můžeme pozorovat rozdělení na čtyři hlavní větve. Odhad, že se v horní polovině grafu vyskytnou dva autoři a ve spodní polovině další dva, se potvrdil jen částečně. V horní polovině se nachází veškeré texty od autorů Kozárová a Neuvedeno, ale mezi nimi se nachází i text od autora Ivan. Ve spodní polovině grafu se nachází ostatní texty od autora Ivan a veškeré texty od autora Čech. Jednotlivé struktury jsou složité a eukleidovská vzdálenost nám v tomto případě neposkytuje příliš jasné výsledky klastrování, na jehož základě není možné dobře identifikovat skupiny textů příslušející jednomu autorovi.



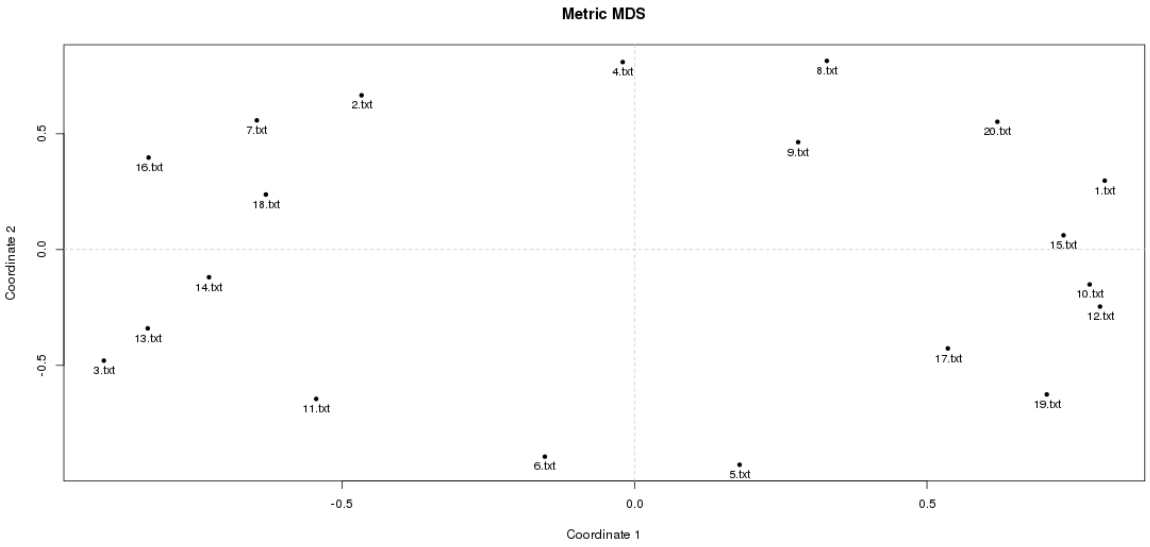
Obrázek 15: Graf výsledků hierarchického shlukování za použití indexů, euklidovská vzdálenost.

Hierarchické shlukování vytvořilo náhled, ve kterém můžeme pozorovat rozdělení na čtyři hlavní větve. Jednotlivé struktury jsou složité. Můžeme odhadovat, že se v horní polovině grafu vyskytnou dva autoři a ve spodní polovině grafu také dva.



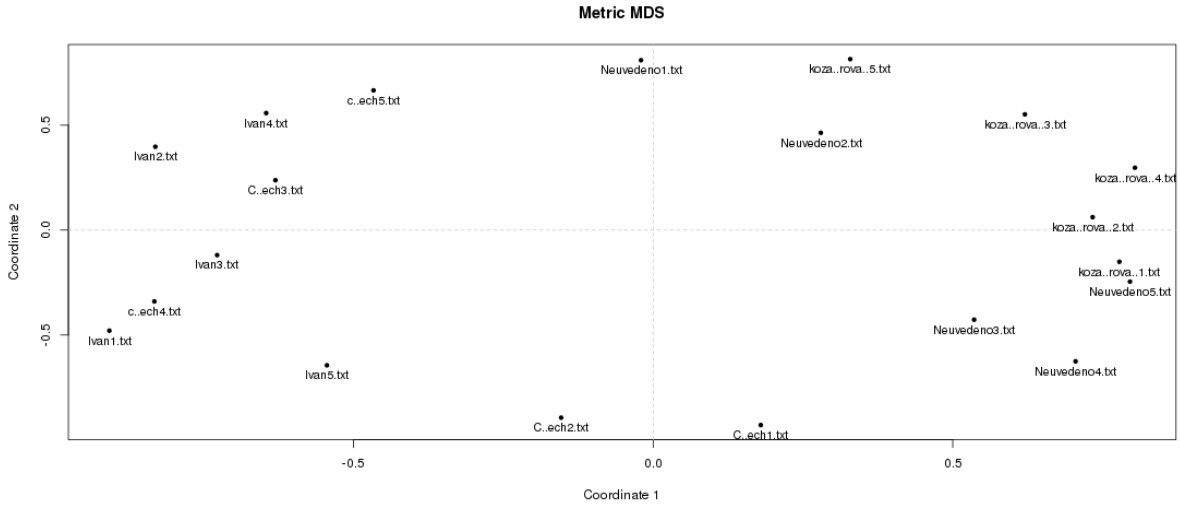
Obrázek 16: Graf výsledků hierarchického shlukování za použití indexů, euklidovská vzdálenost.

Hierarchické shlukování vytvořilo náhled, ve kterém můžeme pozorovat rozdělení na čtyři hlavní větve. Odhad, že se v horní polovině grafu vyskytnou dva autoři a ve spodní polovině další dva, se potvrdil jen částečně. V horní polovině se nachází veškeré texty od autorů Kozárová a Neuvedeno, ale mezi nimi se nachází i text od autora Ivan. Ve spodní polovině grafu se nachází ostatní texty od autora Ivan a veškeré texty od autora Čech. Jednotlivé struktury jsou složité a eukleidovská vzdálenost nám v tomto případě neposkytuje příliš jasné výsledky klastrování, na jehož základě není možné dobře identifikovat skupiny textů příslušející jednomu autorovi.



Obrázek 17: Graf výsledků vícerozměrného škálování za použití indexů, kosinová podobnost.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. Můžeme si všimnout, že se texty rozdělily na dvě poloviny, jednu v pravé části obrázku a druhou v levé části obrázku, daleko od středu. Texty netvoří žádné výrazné shluky a kosinová podobnost nám v tomto případě neposkytuje příliš jasné výsledky shlukování, na jehož základě není možné dobře identifikovat skupiny textů příslušející jednomu autorovi.



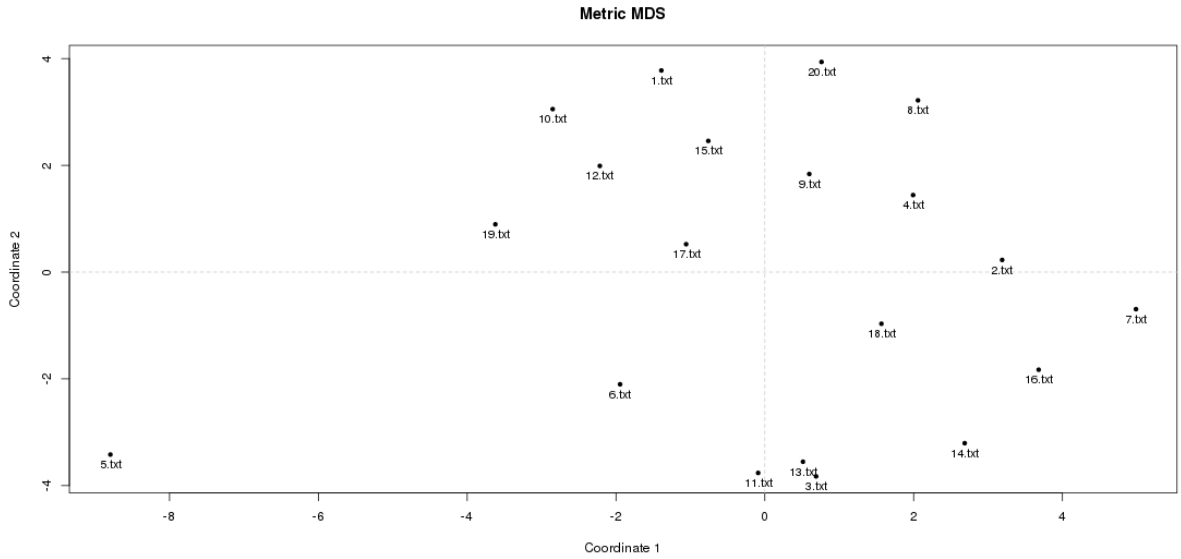
Obrázek 18: Graf výsledků vícerozměrného škálování za použití indexů, kosinová podobnost.

$GOF 0.5755312

V případě této analýzy MDS zrekonstruovalo přibližně 57 % variance vypočítaných vzdáleností.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. Můžeme si všimnout, že se texty rozdělily na dvě poloviny, jednu v pravé části obrázku a druhou v levé části obrázku, daleko od středu. Můžeme pozorovat, že v pravé části grafu se shlukli autoři Kozárová a Neuvedeno a jeden text od autora Čech. V levé části grafu se shlukli autoři Ivan a Čech. Opět není možné autory lineárně oddělit ani sledovat shluky jednotlivých autorů oddělených od sebe.

Vícerozměrné škálování za použití kosinové podobnosti nám nepřináší příliš jasné výsledky, tudíž se potvrzuje naše předchozí domněnka, že není možné odhadovat, které texty by se mohly řadit k jednotlivým autorům.

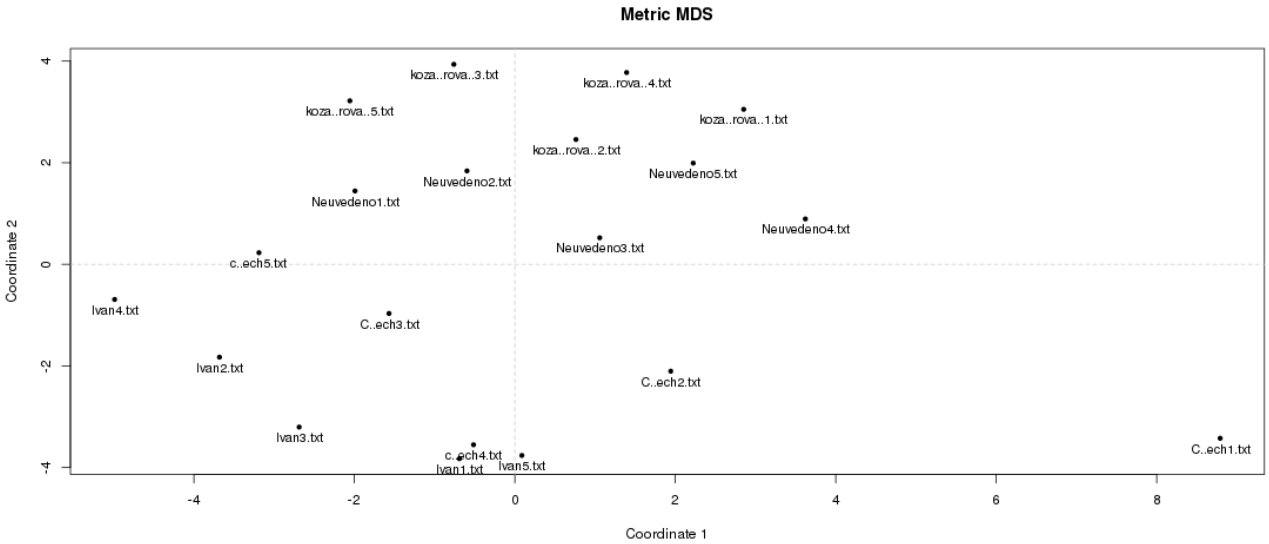


Obrázek 19: Graf výsledků vícerozměrného škálování za použití indexů, euklidovská vzdálenost.

$GOF 0.7599947

V případě této analýzy MDS zrekonstruovalo přibližně 75 % variance vypočítaných vzdáleností.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. Můžeme si všimnout, že opět není možné autory lineárně oddělit, výjimkou je Text 5, který se nachází izolovaný v levém dolním rohu a tvoří tak samostatnou skupinu. Dále by jednu oddělenou skupinu mohly tvořit Texty číslo 3, 11 a 13, které se shlukly v dolní části grafu v těsné blízkosti u sebe.



Obrázek 20: Graf výsledků vícerozměrného škálování za použití indexů, euklidovská vzdálenost.

Vícerozměrným škálováním byl vytvořen náhled na podobnost jednotlivých textů. Potvrdil se náš předchozí odhad, že není možné autory lineárně oddělit, je však možné si všimnout, že všechny texty autora Kozárová se nachází v horní části grafu poměrně blízko sobě navzájem. Pod ní se nachází veškeré texty autora Neuvedeno, také poměrně blízko. Ve spodní části vlevo se nachází všechny texty autora Ivan. Autor Čech má jednotlivé texty nejvzdálenější sobě navzájem, text Čech 1 tvoří v pravé dolní části grafu samostatnou skupinu.

Dále jsme odhadovali, že by jednu oddělenou skupinu mohly tvořit Texty 3, 11 a 13, které se shlukly v dolní části grafu v těsné blízkosti u sebe. Tento odhad se však nepotvrdil, protože texty patří dvěma různým autorům. Vícerozměrné škálování za použití euklidovské vzdálenosti nám v tomto případě nepřináší příliš jasné výsledky.

8.1. Závěr

Analýza pomocí indexů se ukázala jako méně úspěšná než analýza pomocí bag-of-words modelu. Grafy vícerozměrného škálování nám v tomto případě nepřináší jasné výsledky a není možné odhadnout, které texty patří kterým autorům. Za úspěšnější metodu můžeme považovat hierarchické shlukování, stále nám však neposkytuje příliš jasné výsledky klastrování, na jehož základě není možné dobře identifikovat skupiny textů příslušející jednomu autorovi.

8.2. Sumarizace

Cílem této kapitoly bylo obeznámit čtenáře s metodou hodnocení podle vybraných kvantitativně lingvistických indexů. Byly stručně popsány index entropie, index opakování slov a index poměr typů a tokenů. Byla provedena analýza anonymních i neanonymních textů.

9. Závěr praktické části

Cílem praktické části bylo popsat existující metody identifikace autora a navrhnout možnosti, jak doposud využívanou metodiku rozšířit co nejefektivnějším způsobem za využití vícerozměrných metod. Provedli jsme kvantitativně-kvalitativní analýzu, v níž jsme si stanovili kritéria hodnocení anonymních textů a navrhli několik hypotéz o tom, které texty by na základě námi zvolených kritérií mohly patřit k daným autorům, a následně jsme se pokusili za použití kvantitativních metod tyto hypotézy potvrdit nebo vyvrátit. Podařilo se nám potvrdit tři z pěti vyslovených hypotéz. Dále jsme hodnotili, která kritéria se pro naši analýzu ukázala jako užitečná a která naopak nepřinesla žádné využitelné výsledky. Jako neužitečná jsme vyhodnotili následující kritéria: interpunkce, použití velkých a malých písmen a cizích slov, nespisovnost jazyka a komplexita věty. Jako vlastnosti, na jejichž základě je možné rozlišovat autory od sebe navzájem, jsme vyhodnotili tato kritéria: používání emotikonů, dvou a více teček za větou, oslovování ostatních účastníků diskuze, výskyt gramatických chyb a vulgarismů, konfliktnost vyjadřování, rasismus, používání funkce CapsLock a citoslovcí.

V kapitole 6.3. jsme popsali metodu vícerozměrného škálování a vysvětlili pojmy euklidovská vzdálenost a kosinová podobnost.

Interpretací jednotlivých grafů vícerozměrného škálování se nám podařilo oddělit dvě skupiny výrazně odlišných textů – to může svědčit o dvou autorech, nebo o dvou skupinách autorů.

V kapitole 6.4. jsme popsali metoda hierarchického shlukování. Pomocí této metody jsme celkem úspěšně identifikovali dva autory, protože jejich texty tvoří dva samostatné shluky jen s nízkou kontaminací, ostatní autoři jsou buď promíchaní, nebo vzdálení od sebe sama. Při odhalení autorů se nám potvrdilo, že euklidovská vzdálenost nám v tomto případě neposkytuje příliš jasné výsledky klastrování, na jehož základě nebylo možné identifikovat skupiny textů příslušející jednomu autorovi (nebo jedné skupině autorů), jak tomu bylo v případě kosinové vzdálenosti.
V 7. kapitole jsme provedli analýzu pomocí bag-of-words modelu. Tento způsob analýzy se ukázal jako velice úspěšný. Obě výše zmíněné metody se prokázaly jako dobře fungující, hierarchické shlukování však vykazuje o něco lepší výsledky,

protože v grafu od sebe můžeme jasně oddělit všechny čtyři skupiny. Vyskytl se v něm pouze jeden špatně zařazený text.

Cílem 8. kapitoly bylo popsat metodu hodnocení podle vybraných kvantitativně lingvistických indexů. Byla provedena analýza anonymních i neanonymních textů. Analýza pomocí indexů se ukázala jako méně úspěšná než analýza pomocí bag-of-words modelu. Grafy vícerozměrného škálování nám v tomto případě nepřinesly jasné výsledky a nebylo možné odhadnout, které texty patří kterým autorům. Za úspěšnější metodu můžeme považovat hierarchické shlukování, neposkytlo nám však příliš jasné výsledky klastrování, na jehož základě nebylo možné dobře identifikovat skupiny textů příslušejících jednomu autorovi.

Vytyčený cíl praktické části, seznámení čtenáře se způsoby využití vícerozměrných metod a jejich aplikace na dané případy, byl naplněn.

Závěr

Hlavními cíli této bakalářské práce bylo seznámit čtenáře s oborem forenzní lingvistika a představit metody, které se v současnosti používají k identifikaci autora. Doposud využívanou metodiku jsme se následně pokusili co nejefektivnějším způsobem rozšířit o využití vícerozměrných metod.

Práce se skládá ze dvou částí – teoretické a praktické. Teoretická část byla pojata jako seznámení s forenzní lingvistikou jako samostatným oborem a byla rozdělena do 4 kapitol. První kapitola se zaměřovala na obeznámení čtenáře s oborem forenzní lingvistika. Ve druhé a třetí kapitole jsme se věnovali historii nejdříve světové, poté české forenzní lingvistiky a představili jsme si nejvýznamnější osobnosti, díla a také případy, ve kterých byla forenzní lingvistika využita v soudní praxi. Poslední teoretická kapitola hovořila o tom, jaké druhy textů se ve forenzní lingvistice používají, jak se dělí a jaké podmínky musí splňovat, aby mohly být použity k analýze. Cíle teoretické části se nám podařilo naplnit.

Prvním cílem praktické části bylo vytvořit stručný a přehledný souhrn metod, které se v českém prostředí používají k identifikaci autora textů a následně je prakticky aplikovat na námi vybrané anonymní texty. Analýza zkoumaných textů byla koncipovaná tak, aby při tomto průzkumu a hodnocení metod sám autor bakalářské práce neznal odpověď na otázku, které anonymní texty patří ke kterému autorovi. Toho bylo docíleno tak, že texty byly anonymizovány jinou osobou. Texty byly náhodně pojmenovány Text 1 – Text 20. Jako první byla provedena kvantitativně-kvalitativní analýza, ve které jsme stanovili kvalitativní kritéria hodnocení anonymních textů. Jako hodnotící kritéria jsme si zvolili čtrnáct znaků jazykového vyjadřování, které bývají pro jednotlivé autory typické. Těmito kritérii jsou interpunkce, používání velkých a malých písmen, citoslovcí, emotikonů, výskyt dvou a více teček za větou, oslovování, výskyt cizích slov, výskyt gramatických chyb, nespisovného jazyka a vulgarismů. Dále jde o komplexitu vět, konfliktnost, rasismus a projevy nenávisti vůči minoritním skupinám, používání funkce CapsLock nebo jiného zvýraznění slov. Jednotlivá kritéria jsme na základě kvalifikovaného odhadu ohodnotili na škále od 0 do 10, kdy 10 označuje nejvyšší hodnotu a 0 úplně chybí. V podkapitole 6.1. jsme se pokusili navrhnout několik hypotéz a na základě výsledků hodnocení poté posoudit, které texty mohly patřit ke stejným autorům. Dále

jsme zjistili, která kritéria se ukázala pro naši analýzu jako přínosná a která nám naopak neposkytla informace užitečné k rozlišení jednotlivých autorů

Dalším cílem této analytické části bylo rozvinout dosavadní používané metody tak, aby dokázaly najednou pojmout veškerá použitá kritéria. Ukázalo se totiž, že pro analytika je velice těžké pojmout celou tabulku kritérií s jejich výsledky, která má čtrnáct sloupců a dvacet řádků, a získat na ni ucelený pohled. Tato práce tedy zkoumala možnosti, jakým způsobem doposud využívanou metodiku nejefektivněji rozšířit. V 6. kapitole byly detailně popsány vícerozměrné metody, hierarchické shlukování a vícerozměrné škálování, stejně tak i použití euklidovské vzdálenosti a kosinové podobnosti. Ke kvantitativní analýze byl využit lingvistický software QUITA.

V podkapitole 6.4. jsme vícerozměrným škálováním vytvořili dva náhledy na podobnost anonymních textů, jeden za využití euklidovské vzdálenosti a druhý za využití kosinové podobnosti, a pokusili se identifikovat shluky jednotlivých textů podle toho, které z nich by mohly patřit ke stejným autorům. Dále jsme stejným způsobem vytvořili dva náhledy na podobnost neanonymních textů. Interpretací jednotlivých grafů vícerozměrného škálování jsme zjistili, že lze oddělit dvě skupiny výrazně odlišných textů – to může svědčit o dvou autorech, nebo o dvou skupinách autorů. Nepodařilo se nám však identifikovat čtyři od sebe jasně oddělitelné skupiny jednotlivých autorů. Můžeme se domnívat, že se tak stalo z důvodu, že kritéria hodnocení, která jsme zvolili, byla příliš slabá, neodrážela autorství, nýbrž jiné vlastnosti textů, nebo nebylo možné od sebe odlišit jednotlivé autory. Musíme uvažovat i o možnosti, že texty psal ve skutečnosti jediný člověk.

V podkapitole 6.4. jsme pomocí hierarchického shlukování vytvořili dva náhledy na podobnost anonymních textů, jeden za využití euklidovské vzdálenosti a druhý za využití kosinové podobnosti. Pokusili se identifikovat shluky textů na jednotlivých větvích a odhadnout, které z nich by mohly patřit ke stejným autorům. Dále jsme stejným způsobem vytvořili dva náhledy na podobnost neanonymních textů. Při odhalení autorů se nám potvrdilo, že eukleidovská vzdálenost nám v tomto případě neposkytly příliš jasné výsledky klastrování, na jehož základě nebylo možné identifikovat skupiny textů příslušející jednomu autorovi (nebo jedné skupině autorů), jako tomu bylo v případě kosinové podobnosti. Závěrem tedy může být, že ze čtyř provedených analýz anonymních textů na základě

námi zvolených a ohodnocených kritérií se za nejúspěšnější dá považovat metoda hierarchického shlukování za použití kosinové podobnosti.

Po interpretaci výsledků vícerozměrného škálování a hierarchického shlukování jsme provedli hodnocení, zda se nám podařilo potvrdit nebo vyvrátit hypotézy, které jsme vyslovili v kapitole 6.5. Podařila se nám potvrdit hypotéza, že autory lze rozdělit podle toho, jestli ve svých textech používají emotikony. Emotikony používají v komunikaci 2 ze 4 autorů, autorka Kozárová dokonce ve všech svých pěti textech.

Druhou stanovenou hypotézu, zda lze autory od sebe odlišit na základě toho, jestli splňují tři kritéria najednou, tedy jejich texty obsahují emotikony, věty ukončené více než jednou tečkou a zároveň se v nich objevuje oslovování ostatních uživatelů, se nám podařilo potvrdit, jelikož tato tři kritéria splňují dva ze čtyř autorů.

Z třetí hypotézy, kterou se nám podařilo potvrdit, vyplývá, že je možné autory rozdělit na dvě skupiny na základě toho, jestli jejich příspěvky obsahují gramatické chyby.

Při naší analýze se nám nepodařilo potvrdit dvě hypotézy. Jako první jsme vyslovili hypotézu, že texty, které se shlukly na základě kritérií rasismus, vulgarita a konfliktnost, budou patřit dvěma autorům. Druhá hypotéza, kterou se nám nepodařilo potvrdit, byla založena na předpokladu, že veškeré texty obsahující funkci CapsLock budou patřit jednomu autorovi.

V kapitole 7. byl čtenáři představen bag-of-words model a byla provedena analýza anonymních i neanonymních textů. Analýza pomocí bag-of-words modelu se ukázala jako velice úspěšná. Obě použité metody se prokázaly jako dobře fungující, hierarchické shlukování však vykazuje o něco lepší výsledky, protože v grafu od sebe můžeme jasně oddělit všechny čtyři skupiny. Vyskytl se v něm pouze jeden špatně zařazený text.

Poslední metoda, kterou jsme pro naše zkoumání zvolili, bylo hodnocení podle vybraných kvantitativně lingvistických indexů. Byla provedena analýza anonymních i neanonymních textů. Tato metoda se ukázala být méně úspěšnou než analýza pomocí bag-of-words modelu. Grafy vícerozměrného škálování nám v tomto případě nepřinesly jasné výsledky a nebylo možné odhadnout, které texty patří kterým autorům. Za spolehlivější můžeme považovat metodu hierarchického shlukování, přesto nám však neposkytla příliš jasné výsledky klastrování, na jehož

základě nebylo možné dobře identifikovat skupiny textů příslušející jednomu autorovi.

Jedním z cílů bakalářské práce bylo navrhnout, jak současnou metodiku identifikace autora co nejefektivněji rozšířit o využití vícerozměrných metod. V případě použití tabulky vlastních kritérií se ukázalo být velice efektivní využití MDS. Naopak využití indexů odrážejících např. stylistiku se na takto krátkých textech neosvědčilo.

Vzhledem k výsledkům analýz je patrné, že bag-of-words model je z použitých metod nejúspěšnější. Můžeme to přičítat tomu, že texty používané k forenzně-lingvistické analýze jsou psané spontánně a projevují se v nich vlastnosti idiolektu jednotlivého pisatele. Každý člověk totiž používá zaběhlé fráze, které jsou typické pro jeho osobní projev. Vytyčený cíl práce, navržení efektivního způsobu využití vícerozměrných metod, byl tedy naplněn.

Tato práce by mohla být východiskem pro další výzkum zaměřený na využití složitějších metod, které by nám umožnily vyhodnotit důležitost jednotlivých kritérií. Mezi tyto metody patří například rozklad na hlavní komponenty (PCA), dále pak výpočty podmíněných pravděpodobností, korelací a strojově učící se techniky, ze kterých by bylo možné zjistit, která kritéria jsou přínosná a která naopak přínosná nejsou.

Seznam použité literatury

Tištěné zdroje

ČECH, Radek, Ioan-Iovitz POPESCU a Gabriel ALTMANN. *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc: Univerzita Palackého v Olomouci, 2014. Qfwfq. ISBN 978-80-244-4044-6.

FALTÝNEK, Dan, Dalibor PAVLAS, Ondřej VRABEL a Vladimír MATLACH. *Od kvantitativní lingvistiky k neuronovým sítím: Přístupy k analýze textu*. Olomouc: Univerzita Palackého, 2015, kniha je v editaci.

GREPL, Miroslav a Petr KARLÍK. Příruční mluvnice češtiny. 1. vyd. Praha: NLN, Nakladatelství Lidové noviny, 1995, 800 s. ISBN 80-7106-134-4.

GREPL, Miroslav a Petr KARLÍK. Skladby češtiny. Olomouc: Votobia, 1998, 503 s., ISBN 80-7198-281-4.

KARLÍK, Petr, Marek NEKULA a Jana PLESKALOVÁ, ed. *Nový encyklopedický slovník češtiny*. Praha: NLN, Nakladatelství Lidové noviny, 2016. ISBN 978-80-7422-481-2.

MUSILOVÁ, Václava. Co je forenzní lingvistika I. *Čeština doma a ve světě*. Praha: Ústav českého jazyka a teorie komunikace FF UK, 2005, roč 13, 1 - 2, 66 - 70.

MANNING, Christopher D., Prabhakar RAGHAVAN a Hinrich SCHÜTZE. *Introduction to informationretrieval*. New York: Cambridge University Press, 2008. ISBN 05-218-6571-9.

STRAUS, Jiří. *Kriminalistická technika*. 3. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, 2012. ISBN 978-80-7380-409-1.

Internetové zdroje

BLACKWELL, Susan. HistoryofForensicLinguistics.*EncyclopediaofAppliedLinguistics* [online]. 2013, 1-2 [cit. 2019-07-02].

Československo. Vyhláška č. 37 ze dne 20. dubna 1967 k provedení zákona o znalcích a tlumočnících. In: Sbírka zákonů Československé socialistické republiky. 1967, částka 14, s. 130-135. Dostupná také z WWW: https://aplikace.mvcr.cz/sbirka-zakonu/SearchResult.aspx?q=%20o%20znalc%c3%adch%20a%20tlumo%c4%8dn%c3%adc%c3%adch&typeLaw=zakon&what=Text_v_anotaci

Československo. Zákon č. 36 ze dne 20. dubna 1967 o znalcích a tlumočnících. In: Sbírka zákonů Československé socialistické republiky. 1967, částka 14, s. 125 -129. Dostupný také z WWW: https://aplikace.mvcr.cz/sbirka-zakonu/SearchResult.aspx?q=%20o%20znalc%c3%adch%20a%20tlumo%c4%8dn%c3%adc%c3%adch&typeLaw=zakon&what=Text_v_anotaci

HARDAKER, Claire.*The Case of Jenny Nicholl* [online]. [cit. 2019-07-01]. Dostupné z: https://wp.lancs.ac.uk/drclaireh/2012/10/01/the-case-of-jenny-nicholl/

Howcarefulanalysisof text messageshelped police to catch a killer.*Darlington and StocktonTimes* [online]. 7.3.2008 [cit. 2019-07-01]. Dostupné z: https://www.darlingtonandstocktontimes.co.uk/news/2102472.how-careful-analysis-of-text-messages-helped-police-to-catch-a-killer/

International AssociationofForensicLinguistics [online]. [cit. 2019-07-02]. Dostupné z: https://www.iafl.org/about-iafl/

Jenny 'murdered by marriedlover' [online]. 15.1.2008 [cit. 2019-08-13]. Dostupné z: http://news.bbc.co.uk/2/hi/uk_news/england/north_yorkshire/7189805.stm

Ministerstvo spravedlnosti České republiky: Evidence znalců a tlumočníků [online]. [cit. 2019-08-17]. Dostupné z: http://datalot.justice.cz/justice/repznatl.nsf/$$SearchForm?OpenForm

OLSSON, John.*WhatisForensicLinguistics?* [online]. , 4-5 [cit. 2019-08-13]. Dostupné z: https://www.thetext.co.uk/what_is.pdf

ProfessorMalcolmCoulthard [online]. [cit. 2019-07-02]. Dostupné z: https://www2.aston.ac.uk/lss/staff-directory/coulthardm

OWEN, Amos. The text trap.*TheNorthern Echo* [online]. 27.2.2008 [cit. 2019-08-13]. Dostupné z: https://www.thenorthernecho.co.uk/news/2076811.the-text-trap/

Soudní znalci z oboru kriminalistiky [online]. [cit. 2019-07-01]. Dostupné z: https://www.grafickeexpertizy.com/

VONDRÁČEK, Miloslav. Citoslovce a částice — hranice slovního druhu. *Naše řeč* [online]. 1998, **81**(1), 29-37 [cit. 2019-07-20]. Dostupné z: http://nase-rec.ujc.cas.cz/archiv.php?art=7419

Přílohy

Text 1	kozárová 4
Text 2	čech 5
Text 3	Ivan 1
Text 4	Neuvedeno 1
Text 5	Čech 1
Text 6	Čech 2
Text 7	Ivan 4
Text 8	kozárová 5
Text 9	Neuvedeno 2
Text 10	kozárová 1
Text 11	Ivan 5
Text 12	Neuvedeno 5
Text 13	čech 4
Text 14	Ivan 3
Text 15	kozárová 2
Text 16	Ivan 2
Text 17	Neuvedeno 3
Text 18	Čech 3
Text 19	Neuvedeno 4
Text 20	kozárová 3