Filozofická fakulta Univerzity Palackého v Olomouci

Katedra obecné lingvistiky

Verifikace autorství na základě kvantitativní analýzy

magisterská diplomová práce

Autor: Bc. Barbora Šonská

Vedoucí práce: Mgr. Vladimír Matlach, Ph.D.

Olomouc

2023

Prohlášení

Prohlašuji, že jsem magisterskou diplomovou práci „Verifikace autorství na základě kvantitativní analýzy“ vypracovala samostatně a uvedla jsem veškerou použitou literaturu a veškeré použité zdroje.

V Olomouci dne Podpis

Abstrakt

Název práce: Verifikace autorství na základě kvantitativní analýzy

Autor práce: Bc. Barbora Šonská

Vedoucí práce: Mgr. Vladimír Matlach, Ph.D.

Počet stran a znaků: 70 (113065)

Počet příloh: 6

Abstrakt: Problematika ověřování a určování autorství je nahlížena optikou mnoha různých vědních oborů a zkoumána za použití nejrůznějších přístupů. V této práci se zaměříme na ověřování autorství prostřednictvím kvantitativních metod. Pro naše účely jsme vybrali tři z nich: textové ukazatele, model Bag-of-Words a nízkofrekvenční lexikum hapax legomenon. Za pomoci faktorové analýzy vybereme textové ukazatele a jejich výpovědní hodnotu ve vztahu k otázce autorství zkoumaných textů vyhodnotíme prostřednictvím logistické regrese. Dále se zaměříme na použití modelu Bag-of-Words. Výstupy analýz, které získáme použitím tohoto modelu, budeme prezentovat prostřednictvím vizualizační metody vícerozměrného škálování. Třetím zkoumaným přístupem je využití nízkofrekvenčního lexika hapax legomenon jako ukazatele autorského stylu. Zde pro vizualizaci výsledků analýz použijeme nejen metodu vícerozměrného škálování, ale pro porovnání informačního přínosu použijeme i metodu hierarchického shlukování. Tato práce si klade za cíl ukázat, které z vybraných textových ukazatelů mají statisticky významný přínos a na příkladech grafických vizualizací modelu Bag-of-Words, a modelu Bag-of-Words aplikovaného na hapax legomenon, srovnáme jejich informační přínos pro vyhodnocení autorství textu.

Klíčová slova: určování autorství, hapax legomenon, forenzní lingvistika, kvantitativní lingvistika, model Bag-of-Words, vícerozměrné škálování, hierarchické shlukování, tokenizace, indexy

Abstract

Title: Authorship verification based on quantitative analysis

Author: Bc. Barbora Šonská

Supervisor: Mgr. Vladimír Matlach, Ph.D.

Number of pages and characters: 70 (113065)

Number of appendices: 6

Abstract: Authorship verification and attribution are viewed through the lens of many different scientific disciplines and investigated using a variety of approaches. In this work, we focus on authorship verification through quantitative methods. We utilize three methods: text indicators, the Bag-of-Words model, and the low-frequency lexicon hapax legomena. With the help of factor analysis, we select suitable text indicators and evaluate their statistical significance and informative value with respect to the authorship of the examined texts through logistic regression. We then employ the Bag-of-Words model and visualize its results using the multi-dimensional scaling method. Finally, we use the low-frequency lexicon hapax legomena as indicators of author writing styles. In addition to the multi-dimensional scaling method, we visualize its results using the hierarchical clustering method to compare its advantages.

In this work, we determine which of the selected text indicators are statistically significant. We compare the advantages of the Bag-of-Words model and its application to the hapax legomena through graphical visualizations of their results and determine their benefits in the analysis of text authorship.

Keywords: authorship verification, hapax legomenon, forensic linguistics, quantitative linguistics, Bag-of-Words model, multidimensional scaling, hierarchical clustering, tokenization, text indicators

Obsah

1 Úvod....6
2 O autorství....7
2.1 Forenzní lingvistika....8
2.2 Kvantitativní lingvistika....8
2.3 Použitý software....9
2.4 Výběr vstupních dat....9
2.5 Zpracování a příprava vstupních dat....12
3 Textové ukazatele....13
3.1 Faktorová analýza a výběr textových ukazatelů....18
3.2 Logistická regrese....24
3.3 Popis a analýza textových ukazatelů....27
3.3.1 Entropie....27
3.3.2 Type to Token Ratio (TTR)....30
3.3.3 Hirschův index (*h-index*)....32
3.3.4 Délka křivky (*r-index*)....34
3.3.5 Průměrná délka tokenu....36
3.3.6 Výsledky a resumé....38
4 Model Bag-of-Words....39
4.1 Metoda vícerozměrného škálování....40
4.2 Příprava a zpracování vstupních dat....41
4.3 Výsledky a resumé....44
5 Hapax legomenon jako ukazatel autorského stylu....45
5.1 Hierarchické shlukování....45
5.2 Příprava vstupních dat....46
5.3 Zpracování vstupních dat....47
5.4 Výsledky a resumé....52
6 Výsledky analýz....53
7 Diskuze....55
8 Závěr....58
9 Přílohy....60
10 Bibliografie....67

1 Úvod

Jak napovídá již samotný název, tato diplomová práce je zaměřena na problematiku ověřování autorství prostřednictvím kvantitativních metod. Pro naše účely jsme zvolili tři z těchto metod, a to textové ukazatele, model Bag-of-Words a nízkofrekvenční lexikum hapax legomenon. Každé z těchto metod je věnovaná samostatná část, ve které bude krátce představeno její teoretické pozadí, poté bude následovat její praktické užití v experimentu, kdy bude aplikována na specifický vzorek dat, a následné vyhodnocení provedeného experimentu. Cílem této práce je ověřit na souboru vstupních datových podkladů, které byly zvoleny podle předem stanovených kritérií, účinnost těchto tří vybraných kvantitativních metod a porovnat jejich přínos právě při ověřování autorství.

V úvodních kapitolách seznámíme čtenáře s teoretickými východisky naší práce a zasadíme ji do kontextu pojednávané problematiky. Rovněž přestavíme nástroje, které budeme používat při přípravě a zpracování dat, a jejichž prostřednictvím budeme provádět samotné experimenty (2.3 Použitý software). Důležitá bude kapitola, ve které se podrobněji zaměříme na způsob výběru vstupních dat (2.4 Výběr vstupních dat) a jejich přípravě pro praktické využití v našich experimentech (2.5 Zpracování a příprava vstupních dat).

Po teoretickém úvodu bude následovat první ze tří prakticky zaměřených částí této práce. Za použití faktorové analýzy vybereme konkrétní textové ukazatele, kterými se budeme dále zabývat (3 Textové ukazatele). Práce si klade za cíl ověřit a statisticky vyhodnotit přínos jednotlivých vybraných textových ukazatelů. Jejich výpovědní hodnotu ve vztahu k otázce autorství zkoumaných textů vyhodnotíme prostřednictvím logistické regrese.

V další části se zaměříme na použití modelu Bag-of-Words (4 Model Bag-of-Words). Výstupy analýz budeme prezentovat prostřednictvím vizualizační metody vícerozměrného škálování. Na příkladech grafických vizualizací výsledků se pokusíme ukázat fungování tohoto modelu a vyhodnotit jeho informační přínos při rozhodování o autorství analyzovaných textů.

Třetím zkoumaným přístupem bude využití nízkofrekvenčního lexika hapax legomenon jako ukazatele autorského stylu, respektive modelu Bag-of-Words aplikovaného na nízkofrekvenční lexikum hapax legomenon (5 Hapax legomenon jako ukazatel autorského stylu). Zde pro vizualizaci výsledků analýz použijeme nejen metodu vícerozměrného škálování, ale posléze i metodu hierarchického shlukování. Provedeme porovnání informačního přínosu vyhodnocení analýz prostřednictvím těchto dvou modelů a budeme se

snažit rozhodnout, který z těchto přístupů je vhodnější pro námi vybraný zkoumaný typ dat, která jsou zvolená záměrně tak, aby z literárně-vědního hlediska vykazovala co nejvyšší míru zastoupení společných rysů.

Po prakticky zaměřených kapitolách bude následovat prostor pro vyhodnocení výsledků všech provedených analýz v části 7 Diskuze. Pokusíme se přehledně shrnout poznatky plynoucí z této práce a ponecháme otevřenou možnost pro předložení návrhů na vylepšení, případně rozšíření, provedených experimentů.

Zda-li byly naplněny cíle této práce shrneme v jejím samotném závěru.

2 O autorství

Problematikou ověřování a určování autorství se zabývalo mnoho odborníků i laiků, používaly se nejrůznější metody a postupy, a stále se nikomu nepodařilo přijít s jednoduchým a jednoznačným řešením. (Juola 2008; Holmes 1994)

Důvodů je jistě mnoho. Především autorství samotné není nahlíženo jako exaktně měřitelná veličina, ale jako množina jednotlivých prvků, nebo jinak řečeno soubor znaků autorského stylu, které jsou jednotlivě uchopitelné, můžeme je popsat, identifikovat v textu a dále analyzovat, ale doposud nebyl vytvořen komplexní přístup, který by všechny tyto jednotlivé kroky spojoval ve funkční metodu, která by nám umožňovala provádět kompletní analýzu všech těchto prvků zároveň. (Juola 2008)

S rozvojem výpočetní techniky se postupně rozšiřovaly možnosti jejího využití do všech odvětví vědy včetně lingvistiky a etablovala se nová pomezní disciplína, takzvaná komputační lingvistika. Toto propojení matematiky, informatiky a lingvistiky ve spojitosti s dnešní úrovní vyspělosti výpočetní techniky nám umožňuje analyzovat nebývalé množství textů a nacházet nové možnosti jejich zpracování. (Juola 2008) K tomuto účelu jsou vyvíjeny nejrůznější softwary, které dokážou analyzovat více zkoumaných jevů zároveň, kombinovat tyto zkoumané jevy a usnadňují nám hledání nových jazykových jevů, které by se daly označit například jako autorský znak, ať už je to způsob užití diakritiky nebo například výskyt hapax legomen, kterému se, mimo jiné, budeme věnovat v této práci.

2.1 Forenzní lingvistika

Forenzní lingvistika je jedním z interdisciplinárních oborů aplikované lingvistiky, stojícím na pomezí jazykovědy, práva a kriminalistiky. Využívá však poznatky i dalších vědních oborů, jako například sociologie nebo psychologie. (Jurka a Faltýnek 2017) Jedná se tedy o komplexní vědní obor, jehož hlavním úkolem je analýza jazykového materiálu, ke které přistupuje zejména ze dvou úhlů pohledu. Tím prvním je srozumitelnost jazykového projevu a komunikační strategie, tím druhým je identifikace autora, od jeho profilování až po jmenovité určení konkrétní osoby. Podle předmětu analýzy a užité metodologie rozlišujeme jednotlivé podobory forenzní lingvistiky, např. forenzní fonetiku, forenzní stylistiku, forenzní diskurzivní analýzu atd. Charakterizování prostřednictvím jednotné metodologie je vyloučeno vysokou mírou interdisciplinarity tohoto vědního oboru. (Svobodová 1997)

2.2 Kvantitativní lingvistika

Kvantitativní lingvistika je jednou z disciplín obecné lingvistiky, konkrétně odnoží matematické lingvistiky. Zkoumá a popisuje vlastnosti přirozeného jazyka prostřednictvím metod teorie informace, entropie, statistiky a dalších matematických disciplín. Základy české kvantitativní lingvistiky položil Vilém Mathesius ve své práci *O potenciálnosti jevů jazykových* vydané v Praze roku 1911.

Hlavním cílem kvantitativní lingvistiky je formulovat obecné jazykové zákony, z nich vyvozovat hypotézy a ty následně empiricky testovat. Mezi nejznámější a nejvíce používané bezesporu patří Zipfovy zákony a Menzerath-Altmannův zákon. Jak uvádí Čech, Popescu a Altmann ve své práci *Metody kvantitativní analýzy (nejen) básnických textů* (2014), většina zákonů, podle kterých se řídí jazykové chování, se projevují spíše jako tendence, nikoliv jako soubor pravidel, jako je tomu například u fyzikálních zákonů. Proto kvantitativní lingvistika používá jazykové zákony k odhadu pravděpodobnosti výskytu vybraného jazykového jevu za určitých podmínek. (Čech, Popescu a Altmann 2014)

Od počátku 90. let 20. století můžeme sledovat odklon od ryze empirického zaměření kvantitativní lingvistiky směrem k teoretickému. Tuto tendenci můžeme pozorovat například u autorů publikujících v časopisu *Journal of Quantitative Linguistics*, který je oficiálním fórem *Mezinárodní asociace pro kvantitativní lingvistiku*. (Uhlířová 2017)

2.3 Použitý software

Nástroj, který budeme používat pro zpracování dat a vizualizaci výsledků, nese označení *Quantitative Index Text Analyser*, tedy QUITA.[1] Jedná se o analytický on-line nástroj pro práci s textem, který nám umožňuje provádět rozličné kvantitativní analýzy na základě předem zvolených parametrů. Prostřednictvím tohoto nástroje dokážeme o textu získat informace, kterými jsou například entropie, frekvence nebo průměrná délka tokenu. Své využití tento výpočetní software nachází mimo jiné v kvantitativní lingvistice nebo biosémiotice (např. Owsianková, Faltýnek a Kučera 2018).

Dalším nástrojem, který budeme používat, je aplikace pro vyhodnocení logistické regrese dat, která je dostupná z webových stránek katedry obecné lingvistiky Univerzity Palackého v Olomouci.[2] S pomocí tohoto nástroje je ze vstupních dat možné vytvořit model pro binární rozhodování. Nástroj poskytuje doplňující informace o kvalitě modelu v podobě χ^2 a p-hodnot.

2.4 Výběr vstupních dat

Pro náš experiment jsme vybrali dvacet beletristických textů, které byly publikovány pod jmény dvou různých autorů. Prvotním impulzem při formulování hypotézy byly poznatky a postřehy samotných čtenářů níže analyzovaných děl. Díky tomu, že velmi frekventovaným místem pro diskuze jsou v dnešní době internetová fóra, na kterých jsou názory přispěvatelů dohledatelné i mnoho let zpětně, můžeme pozorovat vývoj, jakým fanouškovská základna těchto autorů v průběhu času procházela. V případě prvního autora panovaly pochybnosti o jeho skutečné identitě především z důvodu jeho enormní snahy vyhnout se jakémukoliv kontaktu s fanoušky a obecný odpor k vystupování na veřejnosti. Totožnost druhého autora byla rovněž předmětem mnoha spekulací. Četnost vydávaných děl, mimo jiné, však mezi čtenáři vzbudila podezření, že se za pseudonymem, a zjevně vyfabulovaným životním příběhem, neskrývá pouze jeden spisovatel, ale dokonce skupina autorů.

Hlavním spojujícím prvkem obou autorů je především forma a žánr. Jejich dílo, tedy povídky a romány, se řadí do stejného literárního žánru. Jejich romány na sebe často navazují a tvoří série s různým počtem dílů. Lze tedy předpokládat, že knihy vydávané v těchto sériích ponesou společné rysy a prvky, jako například jména postav, místní názvy, idiolekty postav

1 Dostupné z https://kol.ff.upol.cz/quita
2 Dostupné z http://kol-apps.ff.upol.cz/log-reg

nebo leitmotivy. Dále předpokládáme, že společným rysem textů obou autorů bude žánrově podmíněná stylistika, ať už se bude jednat o vulgarismy, sexismy nebo společensky a politicky nekorektní výrazy. Další žánrově vázané společné prvky, jejichž výskyt můžeme předpokládat, jsou například archetypy postav. Rovněž situace a události, kterým jsou postavy v knihách vystaveny, se dají předvídat. Tato tematická podobnost by měla zapříčinit, že prvky autorského stylu budou splývat, respektive budou těžko odlišitelné od prvků podmíněných obsahem, a díky tomu budeme schopni lépe ověřit citlivost metod, se kterými budeme pracovat.

Z knih vydaných pod jménem prvního spisovatele, pro účely naší práce ho budeme označovat jako *Autora A*, bylo náhodně vybráno 10 titulů, které mají reprezentovat průřez jeho dílem a splňují níže uvedené parametry. Seznam děl uvádíme v *Tabulce 1*.

Z knih vydaných pod jménem druhého spisovatele, pro účely naší práce ho budeme označovat jako *Autora B*, bylo náhodně vybráno taktéž 10 titulů, které mají rovněž reprezentovat průřez jeho dílem a splňují níže uvedené parametry. Seznam děl uvádíme v *Tabulce 2*.

Bibliografické údaje jednotlivých děl nebudou z důvodu ochrany osobních údajů a duševního vlastnictví autorů v této práci uvedeny.

Pro výběr zkoumaných textů jsme stanovili následující čtyři hlavní parametry:

1. **Literární žánr formální** – V našem případě se jedná o beletristické umělecké texty psané formou povídky a románu. Nad rámec tohoto kritéria, ale ku prospěchu našeho účelu, je fakt, že oba autoři vydávají značnou část svých románů v sériích. Vycházíme tudíž z předpokladu, že se tím zvyšuje pravděpodobnost výskytu společných prvků v textech jednotlivých sérií.

2. **Rozsah textu** – Minimální rozsah jednotlivých textů jsme stanovili na 500 tokenů, maximální rozsah není pevně stanoven. Rozsah textu, se kterým budeme pracovat, je 4500 tokenů. To nám umožnilo zařadit do experimentu i povídky, které nejsou součástí sbírky, ale byly publikovány samostatně, bez nutnosti slučovat více krátkých textů do jednoho souboru, čímž by mohlo dojít k ovlivnění experimentu.

3. **Literární žánr obsahový** – V případě metod, které budeme používat, není shodný žánr obsahu analyzovaných textů nutnou podmínkou. My ovšem vycházíme z předpokladu, že pokud jsou texty psané ve stejném žánru, opět se tím zvyšuje pravděpodobnost výskytu shodných prvků, v tomto případě napříč všemi texty

jednotlivého autora. Proto jsme zvolili spisovatele, jejichž díla jsou nejen řazena do stejného literárního žánru, ale dokonce je jeden autor považován za následníka druhého autora.

4. **Původ textu** – Podle informací prezentovaných veřejnosti předpokládáme, že jsou autoři textů stejného pohlaví, rasy i národnosti. Oba jsou to bílí muži ze střední společenské třídy, jsou narozeni ve stejném státě, přibližně ve stejnou dobu, píší stejným jazykem a jsou přibližně stejného věku. Z hlediska literární vědy můžeme oba autory zařadit do stejné epochy.

Předpokládáme, že čím podobnější jsou si jednotlivé texty, tím těžší je odlišit znaky autorského stylu od žánrových a obsahových prvků. Kritéria pro výběr vstupních dat použitých v našem experimentu jsou nastavena tak, aby si texty byly co nejpodobnější s cílem ověřit nejen funkčnost, ale i citlivost používaných metod.

Název díla	**Série**	**Literární žánr**	**Rozsah**
Text 1	Série A	Román	69084 tokenů
Text 2	Série A	Román	58810 tokenů
Text 3	Série A	Román	66688 tokenů
Text 4	Série B	Román	57089 tokenů
Text 5	Série B	Román	65141 tokenů
Text 6	Série C	Povídka	58234 tokenů
Text 7	-	Román	59601 tokenů
Text 8	Série C	Povídka	65794 tokenů
Text 9	Série C	Povídka	15829 tokenů
Text 10	Série A	Román	62323 tokenů

***Tabulka 1**: Seznam náhodně vybraných titulů vydaných pod jménem autora, kterého pro účely naší analýzy budeme označovat jako **Autora A**.*

Název díla	Série	Literární žánr	Rozsah
Text 11	-	Povídka	5757 tokenů
Text 12	-	Povídka	4673 tokenů
Text 13	-	Povídka	11537 tokenů
Text 14	-	Román	73829 tokenů
Text 15	Série D	Román	87821 tokenů
Text 16	Série D	Román	70366 tokenů
Text 17	Série E	Román	98008 tokenů
Text 18	Série E	Román	88904 tokenů
Text 19	-	Román	104052 tokenů
Text 20	-	Román	12248 tokenů

***Tabulka 2**: Seznam náhodně vybraných titulů vydaných pod jménem autora, kterého pro účely naší analýzy budeme označovat jako **Autora B**.*

2.5 Zpracování a příprava vstupních dat

Aby byly výsledky co možná nejpřesnější, a aby se s daty dalo vůbec pracovat, bylo potřeba je nejprve upravit. Vstupní datové sady, tedy autorská díla, jsme nejprve museli převést do formátu podporovaného nástrojem QUITA. Většina textů byla ve výchozím formátu uložena ve formě elektronické knihy (EPUB[3]). Pomocí nástroje pro konverzi formátů jsme díla převedli do formátu PDF[4] a z něj následovně do čistě textové podoby se zřetelem na změnu kódování do znakové sady UTF-8 tak, aby bylo možné s texty dále pracovat v softwaru QUITA (Kubát, Matlach a Čech 2014). Po provedení těchto úprav jsme tak získali k dispozici texty ve formátu srozumitelném pro nástroje, se kterými budeme v rámci našeho experimentu pracovat.

V dalším kroku jsme texty očistili. V našem případě, tedy při práci s beletristickými texty, musíme pro další zpracování výchozí texty upravit tak, aby data nebyla kontaminovaná. Nutné je tudíž odstranění těch částí textu, které není možné připsat samotnému autorovi, respektive které nejsou součástí samotného obsahu zkoumaného díla. Cílem těchto úprav je tedy eliminovat zavádějící části textu, které by mohly negativně ovlivnit výsledky našeho experimentu, a ponechat pouze samotný autorský text. Příkladem takových částí textu je tiráž nebo předmluva psaná vydavatelem.

V rámci příprav jsme také provedli několik sad testovacích výpočtů. Příliš dlouhé texty, o délce 50 000 slov a více, nedokáže software QUITA zpracovat. Tím jsme ověřili jeho

3 Formát souboru je standardizovaný dle https://www.iso.org/standard/63567.html
4 Formát souboru je standardizovaný dle https://www.iso.org/standard/75839.html

výpočetní limity a mohli podle nich stanovit, jaký maximální rozsah mohou mít texty, které budeme dále analyzovat.

3 Textové ukazatele

Výhodou počítačového zpracování textů je beze sporu rychlost zpracování dat a objem informací, se kterými můžeme pracovat. S technologickým vývojem se postupně zvětšuje výpočetní kapacita a snižuje se tak potřebný výpočetní čas pro zpracování informací. Můžeme tak provádět mnohé výpočty v reálném čase, popřípadě s vyšší přesností a nebo s větším objemem informací. Ačkoliv se neustále rozšiřují možnosti využití technologií pro práci s daty, stále je na nás, abychom počítači definovali, která data jsou pro nás zajímavá, důležitá a stojí za další zpracování.

Změna formátu a vyčištění vstupních dat však nejsou jediné úkony související s přípravou zkoumaného textu k dalšímu zpracování. Obvyklý postup je takový, že v první fázi souvislý text rozdělíme na základní jednotky, které budeme dále zkoumat. Realizace těchto jednotek se nazývají tokeny (Cvrček 2017a) a mohou reprezentovat předem libovolně definované celky, jako jsou například jednotlivé věty nebo slova. Tokeny jsou od sebe obvykle oddělené pomocí nealfanumerických znaků. Tento proces se nazývá *tokenizace*. Pro jeho provedení používáme specializovaný počítačový software, tzv. *tokenizér*. (Petkevič 2017) Důležitost procesu tokenizace spočívá v tom, že právě tímto způsobem můžeme výpočetnímu softwaru označit celky, které jsou pro nás zajímavé, a budeme se chtít věnovat jejich dalšímu zkoumání.

V další fázi zpracování textu si musíme určit rozsah, tedy objem dat, se kterým chceme dále pracovat. Pokud pracujeme s více texty, které vzájemně porovnáváme, je nutné sjednotit jejich délku, abychom dosáhli objektivních, nezkreslených výsledků, neboť naměřené hodnoty textových ukazatelů (indexů) se odvíjí mimo jiné právě od délky textu. Můžeme říct, že indexy jsou na délce textu závislé. (Čech, Popescu a Altmann 2014) Pro analýzu tudíž použijeme pouze část původního vzorku přesně definovanou právě počtem tokenů a způsobem jejich výběru, který může být buď náhodný, nebo zachová původní posloupnost tokenů, tak jak se vyskytují v textu od jeho začátku.

Kromě toho, že jsou hodnoty textových ukazatelů závislé na délce textů, jsou často vysoce korelované i vzájemně mezi sebou. (Čech, Popescu a Altmann 2014) To je další skutečnost, kterou musíme brát v potaz, pokud chceme získávat informace o zkoumaných textech právě jejich prostřednictvím.

Mezi základní textové ukazatele, které budeme v práci využívat, patří *entropie*, míra popisující statistickou neuspořádanost, detailně popsaná v sekci (3.3.1 Entropie), *poměr počtu tokenů a typů* (TTR), popsaný v sekci (3.3.2 Type to Token Ratio (TTR)), *h-index*, související s frekvencí výskytu slov a bohatostí použité slovní zásoby, detailněji popsaný v sekci (3.3.3 Hirschův index (h-index)), *r-index*, taktéž související s bohatostí slovní zásoby, diskutovaný v sekci (3.3.4 Délka křivky (r-index)) a konečně *průměrná délka tokenu* (AVGTOKENLEN), popsaná v sekci (3.3.5 Průměrná délka tokenu).

V níže uvedené *Tabulce 3* můžeme vidět přehled dvaceti zkoumaných textů a hodnoty vybraných textových ukazatelů, které byly naměřené při zachování původní délky textů. Na těchto výchozích hodnotách můžeme pro ilustraci vidět závislostní vztah právě mezi délkou textu a naměřenou hodnotou. Dobře patrné je to například u indexu TTR, který je jedním z ukazatelů slovního bohatství textu, a o němž budeme podrobněji mluvit v jedné z následujících kapitol. Na příkladu tohoto textového ukazatele vidíme přímou úměru právě mezi jeho hodnotou a délkou textu. Obdobnou tendenci potom vidíme v různé míře i u ostatních textových ukazatelů.

TEXT	TOKENS	TYPES	TTR	ENTROPIE	H-INDEX	R-INDEX	AVGTOKENLEN
Autor A - Text 1 (série A)	69084	16306	0,24	11,25	77,00	0,88	4,93
Autor A - Text 2 (série A)	58810	14269	0,24	11,11	74,50	0,88	4,88
Autor A - Text 3 (série A)	66688	14653	0,22	11,05	82,00	0,87	4,85
Autor A - Text 4 (série B)	57089	13362	0,23	11,02	73,50	0,88	4,79
Autor A - Text 5 (série B)	65141	13675	0,21	10,91	81,00	0,86	4,74
Autor A - Text 6 (série C)	58234	13953	0,24	11,09	75,00	0,88	4,83
Autor A - Text 7	59601	13879	0,23	10,98	77,00	0,87	4,78
Autor A - Text 8 (série C)	65794	14601	0,22	11,03	79,00	0,85	4,81
Autor A - Text 9 (série C)	15829	4886	0,31	10,30	39,50	0,90	4,74
Autor A - Text 10 (série A)	62323	13543	0,22	10,97	78,00	0,85	4,82
Autor B - Text 11	5757	2905	0,50	10,36	22,00	0,94	5,28
Autor B - Text 12	4673	2309	0,49	9,97	21,00	0,94	4,99
Autor B - Text 13	11537	5173	0,45	10,86	31,00	0,93	5,14
Autor B - Text 14	73829	18561	0,25	11,42	78,00	0,88	4,92
Autor B - Text 15 (série D)	87821	18043	0,21	11,13	87,00	0,84	4,87
Autor B - Text 16 (série D)	70366	15173	0,22	11,00	79,75	0,86	4,83
Autor B - Text 17 (série E)	98008	20881	0,21	11,29	96,50	0,85	4,78
Autor B - Text 18 (série E)	88904	18891	0,21	11,16	93,00	0,86	4,75
Autor B - Text 19	104052	21559	0,21	11,43	100,50	0,86	4,95
Autor B - Text 20	122248	23051	0,19	11,33	109,33	0,83	4,70

***Tabulka 3**: Výchozí hodnoty vybraných indexů u dvaceti zkoumaných textů při zachování jejich původního rozsahu.*

Abychom měli srovnání a mohli sledovat vývoj hodnot, provedli jsme měření těchto vybraných indexů při stanovení jednotné horní hranice počtu tokenů. Tyto hranice jsme stanovili na 500, 1000, 2000, 3000, 4500 a 5000 tokenů. Každé takové měření jsme potom provedli za použití jak náhodného, tak posloupného výběru tokenů z původního textu. Pro ilustraci uvádíme výsledky měření, při kterých jsme použili nejmenší a největší vzorky původních textů, tedy vzorky s rozsahem 500 a 5000 tokenů, které jsme vybrali oběma výše zmíněnými způsoby.

TEXT	TOKENS	TYPES	TTR	ENTROPIE	H-INDEX	R-INDEX	AVGTOKENLEN
Autor A - Text 1 (série A)	500	352	0,70	8,08	5,00	0,95	4,99
Autor A - Text 2 (série A)	500	358	0,72	8,13	6,50	0,97	5,05
Autor A - Text 3 (série A)	500	360	0,72	8,11	6,00	0,96	5,03
Autor A - Text 4 (série B)	500	369	0,74	8,18	7,00	0,97	5,36
Autor A - Text 5 (série B)	500	331	0,66	7,84	6,75	0,94	4,63
Autor A - Text 6 (série C)	500	356	0,71	8,06	6,50	0,96	5,05
Autor A - Text 7	500	354	0,71	8,13	6,00	0,97	5,02
Autor A - Text 8 (série C)	500	347	0,69	8,00	8,00	0,96	4,67
Autor A - Text 9 (série C)	500	355	0,71	8,07	7,00	0,96	4,78
Autor A - Text 10 (série A)	500	348	0,70	8,02	6,00	0,95	4,94
Autor B - Text 11	500	352	0,70	8,06	7,00	0,97	5,07
Autor B - Text 12	500	345	0,69	7,99	6,50	0,95	4,81
Autor B - Text 13	500	376	0,75	8,20	5,00	0,96	5,12
Autor B - Text 14	500	335	0,67	7,99	5,67	0,95	5,08
Autor B - Text 15 (série D)	500	347	0,69	7,97	6,50	0,94	4,92
Autor B - Text 16 (série D)	500	342	0,68	7,97	8,00	0,96	4,76
Autor B - Text 17 (série E)	500	358	0,72	8,06	5,75	0,94	5,04
Autor B - Text 18 (série E)	500	350	0,70	7,93	6,00	0,94	4,92
Autor B - Text 19	500	353	0,71	8,10	6,00	0,96	4,82
Autor B - Text 20	500	308	0,62	7,83	7,00	0,93	4,53

***Tabulka 4**: Naměřené hodnoty vybraných indexů u dvaceti zkoumaných textů při stanovení jejich jednotného rozsahu na 500 po sobě jdoucích tokenů.*

TEXT	TOKENS	TYPES	TTR	ENTROPIE	H-INDEX	R-INDEX	AVGTOKENLEN
Autor A - Text 1 (série A)	500	360	0,72	8,05	6,00	0,95	4,75
Autor A - Text 2 (série A)	500	357	0,71	8,05	7,00	0,97	4,74
Autor A - Text 3 (série A)	500	370	0,74	8,10	7,00	0,95	4,91
Autor A - Text 4 (série B)	500	370	0,74	8,14	7,00	0,97	4,95
Autor A - Text 5 (série B)	500	341	0,68	7,94	7,00	0,95	4,69
Autor A - Text 6 (série C)	500	361	0,72	8,08	6,00	0,96	4,85
Autor A - Text 7	500	356	0,71	7,98	7,00	0,94	4,56
Autor A - Text 8 (série C)	500	348	0,70	7,91	7,00	0,94	4,64

TEXT	TOKENS	TYPES	TTR	ENTROPIE	H-INDEX	R-INDEX	AVGTOKENLEN
Autor A - Text 9 (série C)	500	356	0,71	8,07	6,67	0,96	4,68
Autor A - Text 10 (série A)	500	362	0,72	8,12	6,00	0,95	4,79
Autor B - Text 11	500	381	0,76	8,25	6,00	0,96	5,18
Autor B - Text 12	500	370	0,74	8,09	7,33	0,97	5,07
Autor B - Text 13	500	389	0,78	8,23	5,00	0,95	5,17
Autor B - Text 14	500	371	0,74	8,13	6,50	0,95	4,83
Autor B - Text 15 (série D)	500	362	0,72	8,03	6,60	0,95	4,84
Autor B - Text 16 (série D)	500	366	0,73	8,12	6,00	0,96	4,95
Autor B - Text 17 (série E)	500	369	0,74	8,09	7,00	0,95	4,86
Autor B - Text 18 (série E)	500	348	0,70	7,96	6,00	0,95	4,70
Autor B - Text 19	500	370	0,74	8,17	6,50	0,97	4,93
Autor B - Text 20	500	370	0,74	8,17	6,00	0,96	4,88

***Tabulka 5**: Naměřené hodnoty vybraných indexů u dvaceti zkoumaných textů při stanovení jejich jednotného rozsahu na 500 náhodně vybraných tokenů.*

Jaký vliv má na výsledky měření způsob výběru tokenů, můžeme odvodit porovnáním naměřených hodnot vybraných indexů, které uvádíme v *Tabulce 4* a v *Tabulce 5*. Obecně lze říci, že při náhodném výběru tokenů došlo k nárůstu počtu typů. S tím je spojený nárůst hodnot textového ukazatele TTR. Hodnoty entropie se zvýšily pouze nepatrně, a to zejména u textů *Autora B*. Ke zvýšení došlo i v případě hodnot h-indexu, zatímco r-index zůstal v podstatě beze změny. Průměrná délka tokenu jako jediná z měřených hodnot mírně poklesla.

TEXT	TOKENS	TYPES	TTR	ENTROPIE	H-INDEX	R-INDEX	AVGTOKENLEN
Autor A - Text 1 (série A)	5000	2414	0,48	10,01	22,00	0,94	4,89
Autor A - Text 2 (série A)	5000	2470	0,49	10,07	22,50	0,94	5,10
Autor A - Text 3 (série A)	5000	2350	0,47	9,96	22,50	0,94	4,93
Autor A - Text 4 (série B)	5000	2310	0,46	9,98	23,20	0,95	4,92
Autor A - Text 5 (série B)	5000	2066	0,41	9,67	23,00	0,93	4,68
Autor A - Text 6 (série C)	5000	2321	0,46	9,92	21,33	0,94	4,90
Autor A - Text 7	5000	2156	0,43	9,70	22,50	0,93	4,80
Autor A - Text 8 (série C)	5000	2309	0,46	9,89	20,50	0,93	4,90
Autor A - Text 9 (série C)	5000	2266	0,45	9,87	22,50	0,94	4,86
Autor A - Text 10 (série A)	5000	2221	0,44	9,82	22,50	0,92	4,85
Autor B - Text 11	5000	2626	0,53	10,27	21,00	0,95	5,28
Autor B - Text 12	4673	2309	0,49	9,97	21,00	0,94	4,99
Autor B - Text 13	5000	2636	0,53	10,22	21,00	0,94	5,15
Autor B - Text 14	5000	2338	0,47	9,93	21,50	0,94	4,91
Autor B - Text 15 (série D)	5000	2215	0,44	9,73	21,00	0,93	4,94
Autor B - Text 16 (série D)	5000	2134	0,43	9,73	23,00	0,93	4,74
Autor B - Text 17 (série E)	5000	2305	0,46	10,00	20,00	0,93	5,02
Autor B - Text 18 (série E)	5000	2398	0,48	9,86	20,50	0,93	4,87
Autor B - Text 19	5000	2277	0,46	9,94	20,00	0,94	4,96
Autor B - Text 20	5000	2250	0,45	9,89	22,00	0,92	4,72

***Tabulka 6**: Naměřené hodnoty vybraných indexů u dvaceti zkoumaných textů při stanovení jejich jednotného rozsahu na 5000 po sobě jdoucích tokenů.*

TEXT	TOKENS	TYPES	TTR	ENTROPIE	H-INDEX	R-INDEX	AVGTOKENLEN
Autor A - Text 1 (série A)	5000	2495	0,50	9,99	20,50	0,93	4,90
Autor A - Text 2 (série A)	5000	2454	0,49	9,97	21,00	0,94	4,93
Autor A - Text 3 (série A)	5000	2332	0,47	9,90	22,50	0,93	4,86
Autor A - Text 4 (série B)	5000	2379	0,48	9,90	21,33	0,94	4,80
Autor A - Text 5 (série B)	5000	2315	0,46	9,88	21,50	0,93	4,74
Autor A - Text 6 (série C)	5000	2372	0,47	9,93	21,50	0,94	4,80
Autor A - Text 7	5000	2372	0,47	9,83	22,00	0,93	4,78
Autor A - Text 8 (série C)	5000	2368	0,47	9,86	22,00	0,92	4,75
Autor A - Text 9 (série C)	5000	2205	0,44	9,72	21,25	0,92	4,72
Autor A - Text 10 (série A)	5000	2365	0,47	9,92	21,50	0,93	4,85
Autor B - Text 11	5000	2614	0,52	10,27	20,33	0,95	5,29
Autor B - Text 12	4673	2309	0,49	9,97	21,00	0,94	4,99
Autor B - Text 13	5000	2700	0,54	10,29	20,00	0,94	5,11
Autor B - Text 14	5000	2615	0,52	10,09	20,33	0,95	4,94
Autor B - Text 15 (série D)	5000	2495	0,50	9,97	21,50	0,93	4,94
Autor B - Text 16 (série D)	5000	2442	0,49	9,94	22,00	0,93	4,87
Autor B - Text 17 (série E)	5000	2452	0,49	9,93	21,00	0,93	4,77
Autor B - Text 18 (série E)	5000	2384	0,48	9,87	22,67	0,93	4,70

TEXT	TOKENS	TYPES	TTR	ENTROPIE	H-INDEX	R-INDEX	AVGTOKENLEN
Autor B - Text 19	5000	2560	0,51	10,05	20,67	0,94	4,93
Autor B - Text 20	5000	2527	0,51	10,05	21,50	0,94	4,78

***Tabulka** 7: Naměřené hodnoty vybraných indexů u dvaceti zkoumaných textů při stanovení jejich jednotného rozsahu na 5000 náhodně vybraných tokenů.*

Obdobné porovnání naměřených hodnot vidíme v *Tabulce 6* a v *Tabulce 7*. Tentokrát jsme však zvolili horní hranici délky textů, tedy 5000 tokenů. Stejně jako v předchozím srovnání, i zde shledáváme, že při náhodném výběru tokenů dochází k navýšení počtu typů a v závislosti na tom i ke zvýšení hodnot indexu TTR. Rovněž hodnoty entropie se v průměru mírně zvýšily. Na rozdíl od kratších vzorků dat v rozsahu 500 tokenů, zde dochází ke snížení hladiny naměřených hodnot h-indexu, r-index zůstává v podstatě beze změny a průměrná délka tokenu mírně poklesá, tak jako v předchozím srovnání.

VÝBĚR	TOKENS	TYPES	TTR	ENTROPIE	H-INDEX	R-INDEX	AVGTOKENLEN
Posloupný	500	349,80	0,70	8,04	6,41	0,96	4,93
Náhodný	500	363,85	0,73	8,08	6,48	0,96	4,85
Posloupný	5000	2318,55	0,47	9,92	21,68	0,94	4,92
Náhodný	5000	2437,75	0,49	9,97	21,30	0,93	4,87

***Tabulka 8:** Průměr naměřených hodnot vybraných indexů u dvaceti zkoumaných textů při stanovení jejich jednotného rozsahu na **500** a **5000** tokenů.*

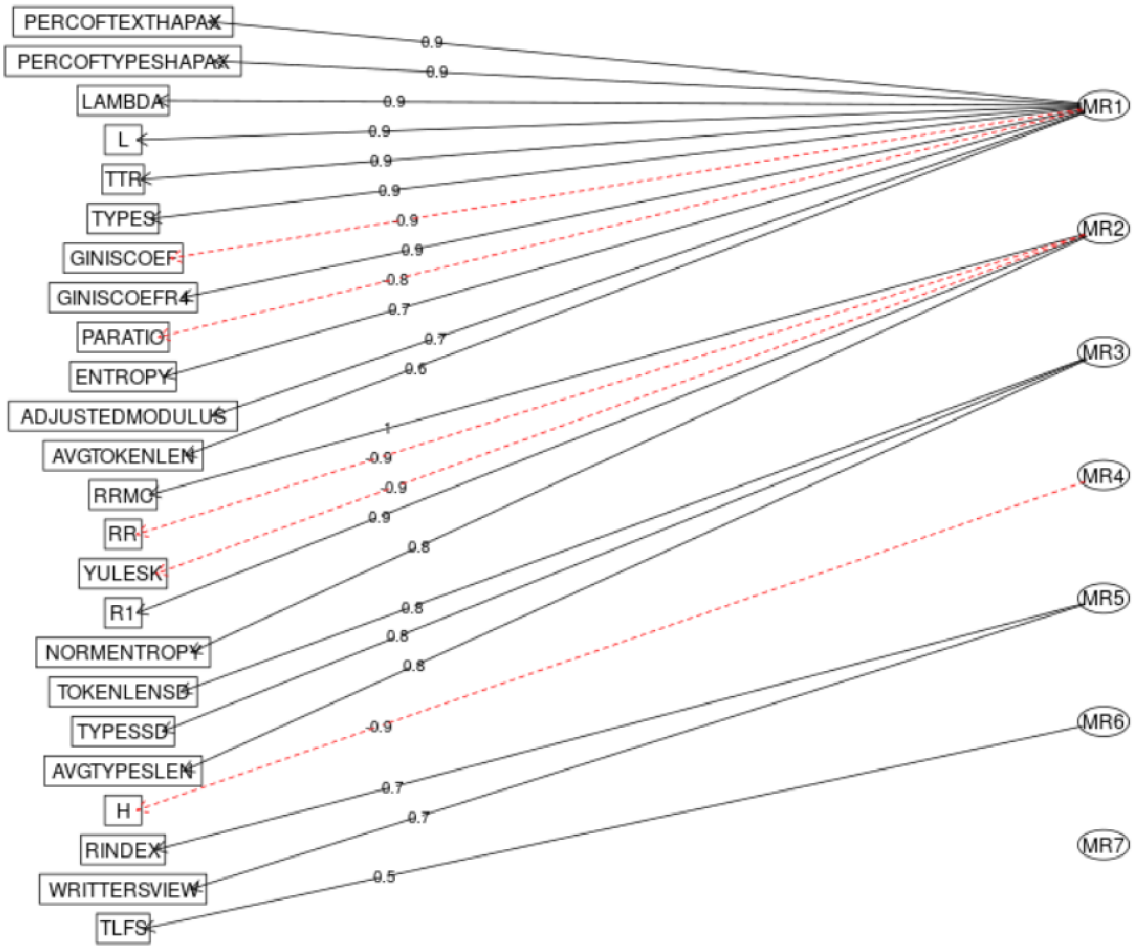
Na podporu výše uvedených tvrzení uvádíme v *Tabulce 8* průměry naměřených hodnot vybraných indexů, kde můžeme zcela jednoznačně vidět, jak velký je nárůst, případně pokles, hodnot ve vztahu k délce zkoumaného vzorku a ke způsobu výběru analyzovaných tokenů.

Na příkladu těchto měření jsme ukázali, jakým způsobem délka textu ovlivňuje výsledné hodnoty sledovaných indexů. Podrobněji se tomuto tématu věnují například Čech, Popescu a Altmann (2014). Nejkratší ze zkoumaných textů, tedy text 12, dosahuje pouze rozsahu 4673 tokenů. Proto jsme rozhodli, že všechny zkoumané vzorky textů v této práci budou mít pevně stanovený rozsah 4500 tokenů, ať už posloupně nebo náhodně vybraných, a to z toho důvodu, aby nebyly negativně ovlivněny výsledky našich měření.

3.1 Faktorová analýza a výběr textových ukazatelů

Faktorová analýza je vícerozměrná statistická metoda, která nám umožňuje zachytit korelační struktury vztahů mezi pozorovanými proměnnými prostřednictvím nalezení společných faktorů a odhalit tak společné vlastnosti pozorovaných proměnných. Tyto společné faktory se v literatuře zpravidla nazývají latentní proměnné. Tato metoda byla původně navržena Charlesem Spearmanem v roce 1904 a byla z počátku používána v oblasti psychologie (Spearman 1904). Později našla široké využití nejen v humanitních a přírodních vědách (např. Malinowski 2002), ale také v marketingové analýze.

Faktorovou analýzu využijeme při výběru textových ukazatelů se kterými budeme dále pracovat. Výpočetní nástroj QUITA podporuje celou řadu textových ukazatelů. Nalezením jejich společných faktorů můžeme vyloučit navzájem korelované textové ukazatele, tedy takové, které využívají obdobných vlastností zkoumaných textů. Na základě informací získaných z nalezené korelační struktury vybereme vhodné ukazatele pro další práci.

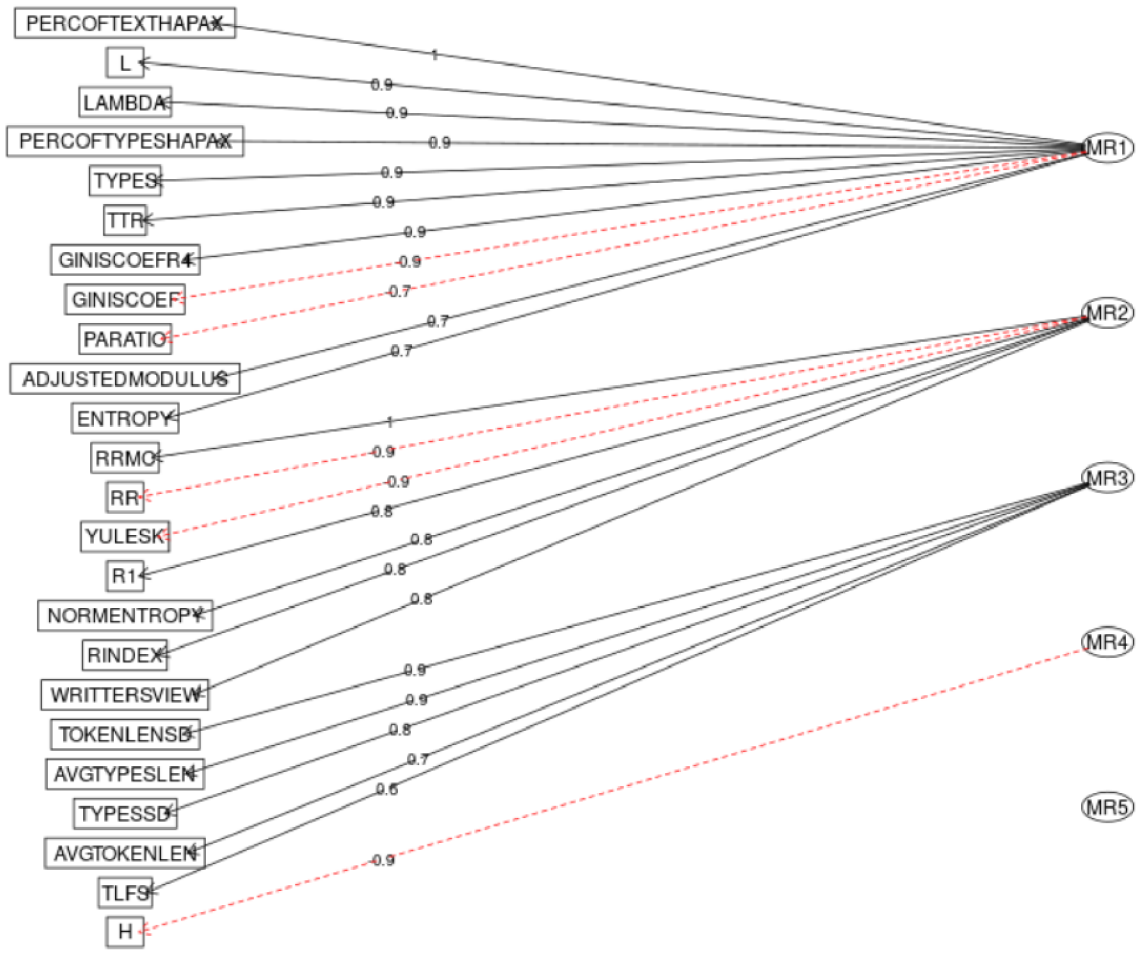


***Obrázek 1**: Faktorová analýza dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 po sobě jdoucích tokenů. V levé části grafu jsou rozmístěny jednotlivé textové ukazatele (pozorované proměnné). V pravé části grafu jsou rozmístěny společné faktory (latentní proměnné). Příslušnost textových ukazatelů ke společným faktorům je vyjádřena čarami. Hodnoty nad čarami vyjadřují spočtenou sílu korelací (hladinu významnosti závislosti). Červené čárkované čáry vyjadřují, že vztah mezi pozorovanou proměnnou a faktorem je nepřímý (takzvaně antikorelovaný).*

S využitím softwaru QUITA jsme provedli faktorovou analýzu dvaceti zkoumaných textů, jejichž délku jsme sjednotili na 4500 tokenů vybraných v tom pořadí, v jakém se vyskytují v původním textu od jeho začátku. Vizualizaci výsledku uvádíme jako *Obrázek 1*. V levém sloupci vidíme výčet všech sledovaných proměnných, tedy všech textových ukazatelů, jejichž hodnoty jsme naměřili. V pravém sloupci vidíme výsledné společné faktory. Spojnice přiřazují jednotlivé latentní proměnné ke sledovaným proměnným a tím znázorňují korelační strukturu mezi proměnnými. Každá z těchto spojnic nese číselnou hodnotu, která reprezentuje spočtenou hladinu významnosti obvykle označovanou jako α. Sledované proměnné jsou

seřazeny nejen podle vazby ke konkrétní latentní proměnné, ale zároveň sestupně podle naměřené hodnoty α. Tím je dotvořena korelační struktura vztahů všech proměnných.

Na *Obrázku 1* dále vidíme, že k faktoru MR7 nejsou přiřazeny žádné ze sledovaných proměnných. Nejedná se ale o defekt provedené faktorové analýzy. Tento zdánlivě nadbytečný faktor ukazuje, že sledované proměnné byly korektně seskupeny k ostatním faktorům a již neexistuje další (potenciálně skrytý) faktor v korelační struktuře, ke kterému by mohly být ještě přiřazeny.



***Obrázek 2**: Faktorová analýza dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 náhodně vybraných tokenů. Detailní popis objektů grafu se nachází u Obrázku 1.*

Následně jsme provedli faktorovou analýzu stejného korpusu dvaceti textů, jejichž délku jsme rovněž sjednotili na 4500 tokenů, s tím rozdílem, že jsme tokeny, které tvoří vzorky jednotlivých textů, vybrali náhodným způsobem. Na *Obrázku 2* vidíme výsledek této analýzy.

Podobně jako u *Obrázku 1* se zde vyskytuje zdánlivě nadbytečný faktor MR5, který ukazuje, že sledované proměnné byly korektně přiřazeny k ostatním faktorům.

Srovnáním výsledků faktorové analýzy na *Obrázku 1* a na *Obrázku 2*, vidíme rozdíly v počtu nalezených latentních proměnných. Dále vidíme rozdílné pořadí některých sledovaných proměnných, nicméně při podrobnějším prozkoumání dojdeme k závěru, že se sledované proměnné shlukují, ve vztahu k latentním proměnným, stejným způsobem jako v případě posloupného výběru tokenů. To by znamenalo, že se korelační struktura v obou analýzách významně podobá, ale v případě náhodného výběru tokenů nám stačí k jejímu zachycení menší počet společných faktorů.

Když se zaměříme na to, které sledované proměnné jsou vázány k jednotlivým společným faktorům, vidíme, že se shlukují podle toho, jaké informace o textu nám přináší. U prvního a druhého společného faktoru, označených MR1 a MR2, se shlukují textové ukazatele slovního bohatství textu. U třetího faktoru (MR3) jsou sdružené indexy vztahující se k délce slov užitých v textu. Čtvrtý faktor (MR4) zahrnuje pouze jeden textový ukazatel, kterým je h-index. V případě posloupného výběru tokenů se část indexů sdružených při náhodném výběru pod MR2 oddělila a vytvořila samostatnou vazbu na faktor MR5. Stejně tak stojí v případě posloupného výběru samostatně s vazbou na MR6 jeden z indexů, který při náhodném výběru náleží k indexu MR3.

Na základě provedených faktorových analýz jsme vybrali 5 textových ukazatelů, kterými jsou TTR, entropie, h-index, r-index a průměrná délka tokenu. Výběr vychází z korelační struktury provedených faktorových analýz a jednotlivé indexy jsou vybrány tak, aby byly zastoupeny všechny společné faktory. Přestože se například TTR a entropie vážou v obou případech ke stejnému společnému faktoru, z povahy těchto indexů a z naměřených hodnot hladiny α jejich významnosti víme, že nám poskytnou o zkoumaných textech různé informace. Proto výběr dvou indexů vázaných na jeden společný faktor není nadbytečný.

Každému z těchto pěti vybraných indexů se budeme podrobněji věnovat v samostatné podkapitole. Zaměříme se na to, jaké informace o textu nám daný index poskytuje, jakým způsobem ho můžeme měřit, a nebo jestli existují korelace s dalšími indexy. Vypočítané hodnoty těchto vybraných textových ukazatelů, se kterými budeme dále pracovat, uvádíme níže v *Tabulce 9* pro posloupný výběr tokenů a v *Tabulce 10* pro náhodný výběr tokenů.

POSLOUPNÝ VÝBĚR TOKENŮ							
TEXT	**TOKENS**	**TYPES**	**TTR**	**ENTROPIE**	**H-INDEX**	**R-INDEX**	**AVGTOKENLEN**
Autor A - Text 1 (série A)	4500	2201	0,49	9,91	20,83	0,94	4,87
Autor A - Text 2 (série A)	4500	2250	0,50	9,98	21,00	0,94	5,09
Autor A - Text 3 (série A)	4500	2189	0,49	9,93	21,00	0,94	4,96
Autor A - Text 4 (série B)	4500	2116	0,47	9,90	21,67	0,95	4,90
Autor A - Text 5 (série B)	4500	1881	0,42	9,60	22,00	0,94	4,67
Autor A - Text 6 (série C)	4500	2148	0,48	9,87	20,00	0,94	4,94
Autor A - Text 7	4500	1975	0,44	9,62	22,00	0,93	4,79
Autor A - Text 8 (série C)	4500	2128	0,47	9,81	19,75	0,93	4,89
Autor A - Text 9 (série C)	4500	2075	0,46	9,79	21,00	0,94	4,87
Autor A - Text 10 (série A)	4500	2055	0,46	9,76	20,00	0,92	4,85
Autor B - Text 11	4500	2394	0,53	10,19	19,50	0,95	5,26
Autor B - Text 12	4500	2254	0,50	9,95	20,00	0,94	5,01
Autor B - Text 13	4500	2434	0,54	10,16	18,60	0,94	5,18
Autor B - Text 14	4500	2132	0,47	9,85	19,50	0,94	4,93
Autor B - Text 15 (série D)	4500	2038	0,45	9,65	20,50	0,93	4,91
Autor B - Text 16 (série D)	4500	1965	0,44	9,65	21,50	0,93	4,72
Autor B - Text 17 (série E)	4500	2109	0,47	9,91	19,00	0,93	5,03
Autor B - Text 18 (série E)	4500	2227	0,49	9,81	18,50	0,93	4,91
Autor B - Text 19	4500	2116	0,47	9,90	18,00	0,94	5,00
Autor B - Text 20	4500	2061	0,46	9,81	20,50	0,92	4,71

***Tabulka 9**: Hodnoty vybraných indexů u dvaceti zkoumaných textů při sjednocení jejich rozsahu na 4500 po sobě jdoucích tokenů.*

NÁHODNÝ VÝBĚR TOKENŮ							
TEXT	**TOKENS**	**TYPES**	**TTR**	**ENTROPIE**	**H-INDEX**	**R-INDEX**	**AVGTOKENLEN**
Autor A - Text 1 (série A)	4500	2313	0,51	9,96	20,00	0,94	4,92
Autor A - Text 2 (série A)	4500	2253	0,50	9,94	21,00	0,94	4,95
Autor A - Text 3 (série A)	4500	2187	0,49	9,87	21,00	0,93	4,87
Autor A - Text 4 (série B)	4500	2236	0,50	9,88	21,33	0,94	4,83
Autor A - Text 5 (série B)	4500	2125	0,47	9,79	19,67	0,94	4,74
Autor A - Text 6 (série C)	4500	2204	0,49	9,86	20,67	0,94	4,82
Autor A - Text 7	4500	2253	0,50	9,91	19,00	0,94	4,91
Autor A - Text 8 (série C)	4500	2246	0,50	9,88	21,33	0,93	4,85
Autor A - Text 9 (série C)	4500	2076	0,46	9,73	19,00	0,93	4,73
Autor A - Text 10 (série A)	4500	2211	0,49	9,85	20,00	0,93	4,84
Autor B - Text 11	4500	2413	0,54	10,19	19,00	0,95	5,29
Autor B - Text 12	4500	2241	0,50	9,94	20,00	0,94	4,99
Autor B - Text 13	4500	2502	0,56	10,23	19,00	0,95	5,12
Autor B - Text 14	4500	2383	0,53	10,01	19,00	0,94	4,91
Autor B - Text 15 (série D)	4500	2241	0,50	9,82	21,00	0,93	4,87
Autor B - Text 16 (série D)	4500	2241	0,50	9,82	21,33	0,94	4,86
Autor B - Text 17 (série E)	4500	2284	0,51	9,93	19,00	0,94	4,79
Autor B - Text 18 (série E)	4500	2181	0,48	9,78	21,00	0,93	4,74
Autor B - Text 19	4500	2317	0,51	9,95	20,50	0,94	4,94
Autor B - Text 20	4500	2263	0,50	9,91	21,33	0,94	4,74

Tabulka 10: *Hodnoty vybraných indexů u dvaceti zkoumaných textů při sjednocení jejich rozsahu na 4500 náhodně vybraných tokenů.*

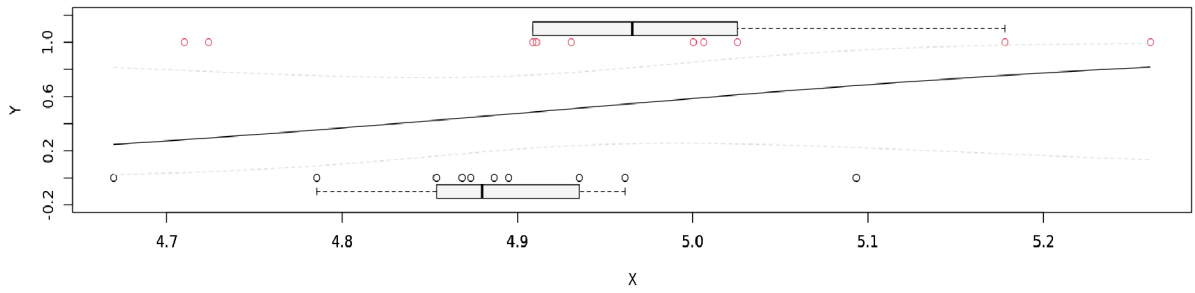
3.2 Logistická regrese

Počátky vývoje logistické regrese sahají až do 19. století, kdy byla pro popis růstu počtu obyvatel použita takzvaná logistická funkce. (Cramer 2002) Metoda logistické regrese se dále rozvíjela až do dnešní podoby a jako jedna z metod matematické statistiky je dnes využívána v celé řadě odvětví při zpracování dat a v rozhodovacích procesech. Příkladem může být bankovní sektor, ve kterém se logistická regrese uplatňuje při procesu schvalování úvěru k vyhodnocení schopnosti klienta danou půjčku splácet (Crook, Edelman a Thomas 2007). V medicíně se logistický model používá například k predikci závažnosti průběhu onemocnění s ohledem na výchozí zdravotní stav pacienta (Dreiseitl a Ohno-Machado 2002). Své využití nalezl také v marketingu, kde lze jeho prostřednictvím mimo jiné předvídat přechod stávajícího klienta ke konkurenci nebo výběr produktu, který klient upřednostní (viz např. Constantin 2015). Obecně lze říci, že model logistické regrese slouží k odhadu pravděpodobnosti výskytu nějakého jevu, který je učiněn na základě známých skutečností, které mohou výskyt tohoto jevu ovlivnit. Více o problematice logistické regrese a jejích aplikacích nalezneme například v knize *Applied Logistic Regression* (Hosmer a Lemeshow 2000).

alpha

0.01

X	Y
4.868444	0
5.093333	0
4.961333	0
4.895111	0
4.669778	0
4.935333	0
4.785556	0
4.886667	0
4.873333	0
4.853778	0
5.261111	1
5.006222	1
5.178222	1
4.930667	1
4.908889	1
4.723778	1
5.025556	1
4.910889	1
5.000444	1
4.710000	1



ExpOddsRatio	Estimate	Std. Error	z value	Pr(>|z|)
0.00	-21.75	17.60	-1.24	0.22
82.92	4.42	3.58	1.24	0.22

```
[!] Model NENÍ s dodaným regresorem významně lepší než bez něj (nemá na určení odpovědí významný vliv).

Chí-kvadrát = 1.754(1), p-hodnota ≈ 0.185382

Navýšení o 1 jednotku v 'x' změní šanci odpovědi '1' 82.922 krát,
... neboli navýšení o 1 jednotku v 'x' zvýší šanci být odpovědí '1' o 8192.2 %.

Logistická regrese dokáže správně přiřadit '0' a '1' v 65 % případů.
```

***Obrázek 3**: Ukázka výstupu z aplikace logistické regrese. Dostupná z http://kol-apps.ff.upol.cz/log-reg Vstupem jsou nezávislé proměnné (sloupec X) a závislé proměnné (sloupec Y) nabývající hodnot 0 a 1. V našem případě tyto hodnoty reprezentují autorství, tedy zda-li je autorem díla **Autor A** či **Autor B**. Výstupem je graf s jednotlivými body, nalezenou logistickou křivkou a vypočtenými parametry. V šedém rámečku je uvedeno automatizované vyhodnocení statistické významnosti (signifikance) výsledného regresního modelu.*

Jak jsme již zmínili v kapitole 2.3 Použitý software, logistickou regresi budeme provádět za použití on-line nástroje, jehož uživatelské rozhraní a vzorové výstupy jsou pro ilustraci uvedeny na *Obrázku 3*. Zpracovávat budeme data uvedená v *Tabulce 9* a *Tabulce 10*. Tato data budeme vkládat do levého sloupce aplikace logistické regrese, který je určený pro *nezávislé* proměnné, takzvané *prediktory*. Ty mohou nabývat libovolných hodnot, v našem případě budeme zpracovávat naměřené hodnoty jednotlivých textových ukazatelů. Ve vizualizaci výsledků budou tyto hodnoty vykresleny na ose X, tedy na horizontální ose grafu. Do pravého sloupce aplikace logistické regrese budeme vkládat binární *závislé* proměnné, které nabývají hodnoty 0 nebo 1. Tyto hodnoty odpovídají výskytu (nebo nevýskytu) sledovaného jevu. V našem případě je tím jevem původce textu, tedy *Autor A* nebo *Autor B*. Přidělení těchto hodnot je arbitrární, tudíž nemá žádný vliv na výsledky. Jediným praktickým důsledkem je barevné rozlišení dat při jejich vizualizaci, kdy data se závisle proměnnou, nabývající hodnoty 0, jsou vykreslena černou barvou a data se závisle proměnnou, nabývající hodnoty 1, jsou vykreslena červeně. Tyto hodnoty nalezneme na ose Y, tedy na vertikální ose grafu. S pomocí logistické regrese proložíme datové body spojitou křivkou, kterou nazýváme *sigmoida*. Kromě vizualizace výsledků je výstupem aplikace logistické regrese i vyhodnocení použitého modelu s procentní účinností. Důležitou sledovanou hodnotou pro nás bude takzvaná *p*-hodnota, která nám říká, jestli máme dostatečnou evidenci, abychom mohli rozhodnout, jestli je výsledek statisticky signifikantní. Abychom mohli výsledný model vyhodnotit právě jako statisticky signifikantní, nesmí *p*-hodnota přesáhnout 0,01.

3.3 Popis a analýza textových ukazatelů

Pro každý z vybraných indexů provedeme logistickou regresi, abychom zjistili, s jakou pravděpodobností dokážeme na základě jejich naměřených hodnot přiřadit texty k jednotlivým autorům. Pokud budou výsledné hodnoty sledovaných indexů signifikantně rozdílné, měli bychom být schopni za pomoci tohoto modelu určit, který z textů náleží *Autorovi A*, a který z textů je dílem *Autora B*.

3.3.1 Entropie

Jedním z významných textových ukazatelů, na který se v naší práci zaměříme, je entropie. Protože se jedná o veličinu přítomnou a důležitou v celé řadě vědních oborů, můžeme se setkat s jejími různými definicemi. Obecně ale můžeme říci, že se jedná o takzvanou „míru neurčitosti“ nebo „míru neuspořádanosti“ systému. Poprvé byl tento pojem zaveden

v klasické termodynamice roku 1865 Rudolfem Clausiem ve vztahu k nevratnosti termodynamických jevů v podobě druhého termodynamického zákona. V teorii informace, u jejíhož zrodu stál Claude Elwood Shannon, byl pojem entropie poprvé zaveden v kontextu náhodných veličin ke kvantifikaci jejich náhodnosti v článku *The Mathematical Theory of Communication* (Shannon 1948). Z této vědní disciplíny byl pojem entropie posléze přenesen do lingvistiky (Holmes 1994).

Entropii statistického souboru, například daného textu, můžeme spočítat s použitím vzorce

$$H = -\sum_{i=1}^{N} p_i \log_x p_i,$$

kde p_i vyjadřuje pravděpodobnost (frekvenci) výskytu *i*tého prvku souboru a N celkový počet prvků souboru. (Cover 2012; Čech, Popescu a Altmann 2014). Základ logaritmu souvisí s volbou jednotek množství informace. V kontextu teorie informace a kvantitativní lingvistiky se zpravidla používá základ logaritmu dva (Čech, Popescu a Altmann 2014), množství informace pak uvádíme v bitech (Shannon 1948). Volba základu nicméně nemá vliv na obecnou informační hodnotu entropie.

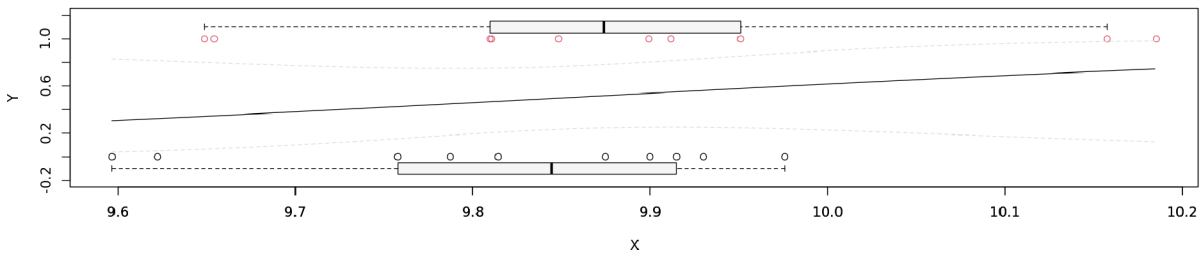
Kromě výše definované informační entropie existují další entropické veličiny používané nejen v teorii informace. Jedná se například o křížovou entropii, podmíněnou a sdruženou entropii (Cover a Thomas 2012), vzájemnou informaci (Shannon 1948; Fano 1957) a nebo relativní entropii (Manning a Schütze 1999), v literatuře také označovanou jako Kullback–Leiblerova divergence a mimo jiné používanou také v lingvistice při určování autorství (Zhao 2006).

V lingvistice entropie přirozeně vyjadřuje redundanci, potažmo bohatost, použitého jazyka, jelikož souvisí s frekvencí výskytu slov a jejich opakováním. (Doležel 1963; Krámský 1959) Čím vyšší je vypočtená hodnota entropie, tím méně koncentrovaný je slovník a tím větší je bohatství textu, neboť jednotlivá slova nesou vyšší množství informace. Z Gibbsovy nerovnosti (Brémaud 1988) vyplývá, že text ve kterém se neopakují slova nabývá nejvyšší možné hodnoty entropie, rovné $\log N$ kde N odpovídá délce textu. Naopak, nižší entropie souvisí s vyšší redundancí textu, tedy s opakováním slov. Uvažujme příklad plně redundantního textu, tedy takového textu, ve kterém se stále opakuje jen jedno jediné slovo. Entropie takového textu je pak nulová, neboť $\log p_i = \log 1 = 0$.

V obecné rovině lze říci, že redundance je přirozenou a potřebnou součástí jazyka, protože chrání sdělení před zkreslením nebo ztrátou informace vlivem šumu. V praxi se může jednat o

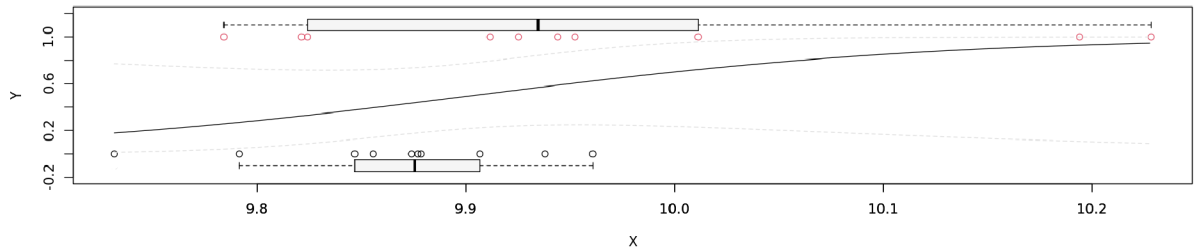
tiskové chyby, překlepy nebo o poškození média, například v důsledku znečištění. V případě dálkové komunikace mohou sdělení ovlivnit přenosové chyby. Informační redundance tak snižuje pravděpodobnost, že přenesená zpráva bude pro příjemce nesrozumitelná. (Shannon 1948)

Při určování autorství je pro porovnání autorských děl možné využít entropie daných textů (Holmes 1994; Kubát, Matlach, Čech 2014), respektive relativní entropie (Zhao 2006). V této práci se omezujeme na použití základního konceptu informační entropie, výše definované symbolem H, vypočítané pro jednotlivá autorská díla nástrojem QUITA. Z vypočtených hodnot užitím logistické regrese vytváříme logistický regresní model. Při analýze výsledků se zaměříme na statistickou významnost regresního modelu. Model budeme považovat za statisticky nevýznamný v případě, že vypočtená p-hodnota přesáhne námi stanovenou hranici 0,01. V takovém případě nebude mít výsledný model v kontextu této práce přínos pro určení autorství.



***Obrázek 4**: Logistická regrese indexu entropie dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 po sobě jdoucích tokenů.*

Na *Obrázku 4* vidíme grafické znázornění výsledků logistické regrese indexu entropie, jehož hodnoty byly naměřeny ve vzorcích zkoumaných textů o jednotné délce 4500 tokenů, které jsme vybrali v jejich původním pořadí od začátku textu. S pomocí získaného modelu můžeme autory k těmto zkoumaným textům správně přiřadit pouze v 50 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí p-hodnoty 0,29566, což znamená, že tento model není funkční a výsledek není statisticky signifikantní. Index entropie při použití posloupného výběru tokenů tudíž není vhodnou metrikou pro rozlišení autorství námi zkoumaných textů.



***Obrázek 5**: Logistická regrese indexu entropie dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 náhodně vybraných tokenů.*

Obrázek 5 nám poskytuje náhled na grafické znázornění výsledku logistické regrese aplikované rovněž na index entropie. V tomto případě jsme však hodnoty indexu měřili na základě náhodného výběru tokenů. Takto získaný model nám umožňuje správně přiřadit autory k textům v 70 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí *p*-hodnoty 0,060587, což znamená, že tento model není funkční, respektive statisticky signifikantní, neboť dostatečnou evidenci pro rozhodnutí jsme stanovili na 0,01. Ani při náhodném výběru tokenů není index entropie vhodnou metrikou pro rozlišení autorství námi zkoumaných textů.

3.3.2 Type to Token Ratio (TTR)

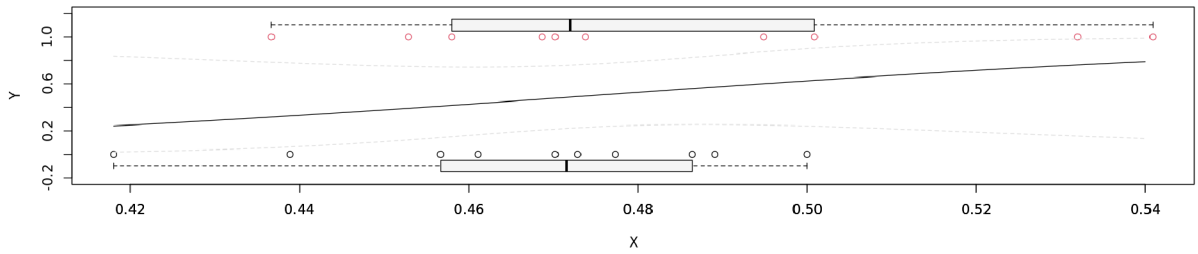
Jak již bylo řečeno, token je každá jedna konkrétní realizace jednotky v daném kontextu. Jejím na kontextu nezávislým protějškem je tzv. *type*. Typy jsou obvykle v základním slovníkovém tvaru, tzv. *lemmata*, a jejich výčet z daného textu tvoří množinu všech tokenů bez jejich opakování. Type rovněž může nést různé kvantifikovatelné vlastnosti, jako například frekvenci. Proto jsou tyto jednotky používány mimo jiné i při sestavování frekvenčních slovníků, které jsou při textových analýzách využívány již od konce 19. století. Příkladem je *Häufigkeitswörterbuch der deutschen Sprache*, sestavený Friedrichem Wilhelmem Kädingem, vydaný v roce 1897.

Pokud se podíváme, v jakém poměru se v textu vyskytují typy a tokeny, dokážeme určit míru lexikální rozmanitosti zkoumaného textu. Tento ukazatel, tedy *type to token ratio*, zkráceně TTR, se vypočítá tak, že počet unikátních slov (typů) vydělíme počtem všech slov realizovaných v daném textu (tokenů). Formální matematický zápis potom vypadá následovně.

$$TTR = \frac{V}{N},$$

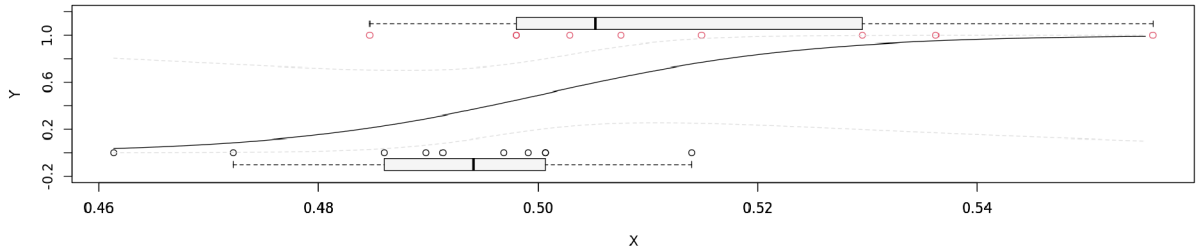
kde počet tokenů označujeme písmenem N, počet typů potom zastupuje písmeno V, referující k anglickému výrazu pro slovník, *vocabulary*. (Čech a Kubát 2017) Výsledné hodnoty se pohybují v rozpětí 0 až 1, přičemž při výsledné hodnotě 1 je každé slovo v textu unikátní, text se tedy vyznačuje nejvyšší možnou mírou entropie. Čím více se hodnota TTR blíží nule, tím více se slova opakují a tím narůstá míra redundance textu (Sedlačíková 2012). Tento ukazatel tedy svým způsobem doplňuje entropii (popsanou podrobněji v sekci 3.3.1 Entropie). Na tuto skutečnost poukazuje také faktorová analýza provedená v úvodu této kapitoly (3.1 Faktorová analýza a výběr textových ukazatelů), kde se jak na *Obrázku 1*, tak na *Obrázku 2*, ukazatele TTR a entropie sdružují u stejného faktoru.

V této sekci se pokusíme ověřit, zda-li je možné ukazatel TTR využít k přiřazení (ověření) autorství. Nejprve s pomocí programu QUITA vypočítáme hodnoty TTR pro jednotlivé texty a na základě těchto hodnot logistickou regresí vytvoříme rozhodovací model. Při analýze se zaměříme na statistickou významnost modelu se zvolenou hranicí významnosti 0,01.



***Obrázek 6**: Logistická regrese indexu TTR dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 po sobě jdoucích tokenů.*

I v případě indexu TTR jsme nejprve měřili jeho hodnoty na základě posloupného výběru tokenů. Výsledný graf logistické regrese vidíme na *Obrázku 6.* Získaný model nám umožňuje správně přiřadit autory k textům v 50 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí *p*-hodnoty 0,216063, což znamená, že tento model není funkční, respektive statisticky signifikantní. Dostatečnou evidenci pro rozhodnutí jsme stanovili na 0,01. Tento index tudíž není vhodnou metrikou pro rozlišení autorství textů zkoumaných v této práci.



***Obrázek** 7: Logistická regrese indexu TTR dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 náhodně vybraných tokenů.*

Pro srovnání jsem opět provedli měření hodnot indexu TTR i při náhodném výběru tokenů. Výsledný graf vidíme na *Obrázku 7*. S pomocí takto získaného modelu můžeme autory k textům správně přiřadit v 65 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí *p*-hodnoty 0,009773, což znamená, že tento model je funkční a přináší nám statisticky signifikantní výsledek. Index TTR je tudíž při náhodném výběru tokenů vhodnou metrikou pro rozlišení autorství námi analyzovaných textů.

3.3.3 Hirschův index (*h-index*)

Hirschův index, zpravidla označovaný jako *h-index*, popřípadě jako h-bod, byl navržen Jorge E. Hirschem pro účely charakterizace publikační činnosti vědeckých pracovníků fyzikálních disciplín (Hirsch 2005). Následně se stal jedním z nejčastěji používaných indikátorů ve scientometrii (Quaia a Vernuccio 2022). Svůj význam a použití našel také na poli kvantitativní lingvistiky (Popescu a Altmann 2006). S jeho pomocí můžeme například rozdělit takzvaný frekvenční slovník na autosémantickou a synsémantickou část, charakterizovat slovní bohatství textu, jeho tématickou koncentraci a nebo kompaktnost textu (Popescu 2007; Popescu 2009; Čech 2016).

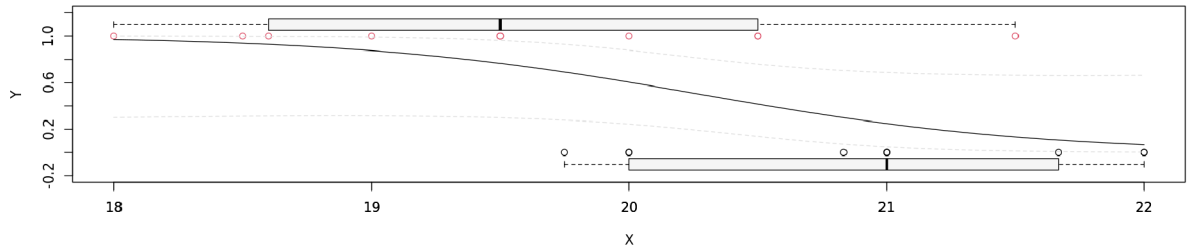
Hodnotu h-indexu můžeme určit ze slovníku daného textu tak, že jednotlivá slova slovníku seřadíme *sestupně* podle jejich frekvence (zastoupení) v textu. Funkce $f(r)$ udává frekvenci *r*tého nejčastějšího slova. Pořadí r zpravidla označujeme termínem *rank*. Pevný bod této funkce, tedy takové h, pro které platí $f(h)=h$, určuje h-index. Neboli, slovo, jehož pořadí (rank) se rovná jeho frekvenci, určuje h-index, respektive hodnota h-indexu je rovna pořadí tohoto slova.

V případě, že takové slovo neexistuje, je možné využít složitějšího výpočtu, založeného na vyřešení soustavy dvou rovnic, a výsledek zaokrouhlit na celé číslo. Výpočet spočívá v

$$h=\left[\frac{f(r_i)r_j-f(r_j)r_i}{r_j-f(r_j)-r_i+f(r_i)}\right],$$

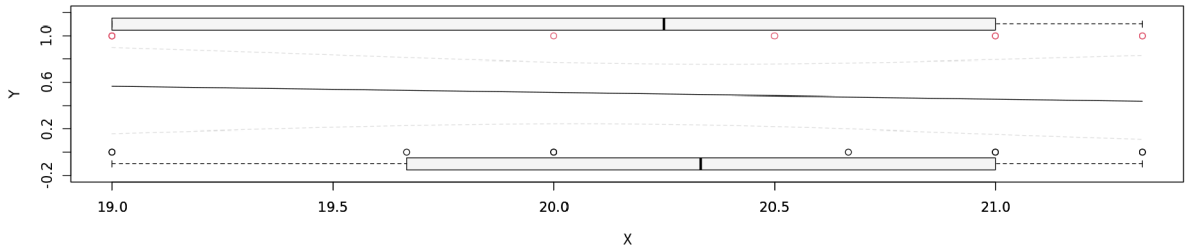
kde při výpočtu pracujeme s ranky r_i a r_j pro které platí, že $r_i<r_j$ a r_i je takové největší číslo, pro které platí $r_i<f(r_i)$ a obdobně, r_j je takové největší číslo, pro které platí $r_j<f(r_j)$. Neboť výsledkem tohoto výpočtu je obecně reálné číslo, musíme jej pro určení celočíselného h-indexu zaokrouhlit. Tuto úpravu ve výrazu symbolizují hranaté závorky (Popescu 2009).

Jelikož je h-index spjatý s celou řadou lingvistických charakteristik zkoumaného textu (viz výše), pokusíme se ověřit, zda-li je možné jej přímo využít k přiřazení (ověření) autorství. Nejprve s pomocí programu QUITA vypočítáme hodnoty h-indexu pro jednotlivá díla a z těchto hodnot pak logistickou regresí vytvoříme rozhodovací model. Při analýze výsledného modelu se opět zaměříme na jeho statistickou významnost s hranicí významnosti 0,01.



***Obrázek 8**: Logistická regrese h-indexu dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 po sobě jdoucích tokenů.*

Nejdříve jsme použily hodnoty h-indexu zkoumaných textů naměřených při posloupném výběru tokenů. Na *Obrázku 8* vidíme grafické vyobrazení takto získaného modelu. Na jeho základě můžeme k textům správně přiřadit autory v 70 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí *p*-hodnoty 0,003505. Protože jsme jako hranici významnosti stanovili hodnotu 0,01, znamená to, že tento model je funkční a výsledek je statisticky signifikantní. H-index při použití posloupně vybraných tokenů je tedy vhodnou metrikou pro rozlišení autorství námi zkoumaných textů.



***Obrázek 9**: Logistická regrese h-indexu dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 náhodně vybraných tokenů.*

Při náhodném výběru tokenů můžeme s pomocí získaného modelu logistické regrese pro h-index správně přiřadit autory k textům v 50 % případů, viz *Obrázek 9*. Kvalitu tohoto modelu můžeme charakterizovat pomocí *p*-hodnoty 0,656651, což znamená, že tento model není funkční, respektive statisticky signifikantní. Dostatečnou evidenci pro rozhodnutí jsme stanovili na 0,01, stejně jako v předešlých případech. H-index naměřený na vzorku náhodně vybraných tokenů proto není vhodnou metrikou pro rozlišení autorství textů, které zkoumáme v této práci.

3.3.4 Délka křivky (*r-index*)

Poměrná délka křivky, takzvaný *r-index*, úzce souvisí se slovním bohatstvím textu (Čech, Popescu a Altmann 2014). Pro její výpočet zpočátku postupujeme podobně jako v případě h-indexu. Nejprve setřídíme slovník daného textu *sestupně* podle frekvence výskytu jednotlivých slov. Funkce $f(r)$ pak udává frekvenci *r*tého nejčastějšího slova.

Délku křivky této funkce, zpravidla označovanou symbolem L, můžeme spočítat z délek úseček (segmentů) mezi jednotlivými funkčními hodnotami $f(1), f(2), f(3), \ldots, f(V)$ kde V udává celkový počet slov slovníku. Délku jednotlivých segmentů určíme s pomocí Pythagorovy věty

$$d(r)=\sqrt{(f(r+1)-f(r))^2+1},$$

kde symbolem $d(r)$ rozumíme vzdálenost mezi prvním a druhým bodem frekvenční funkce. Tento výraz se v literatuře také označuje jako Euklidovská vzdálenost. Celková délka křivky pak odpovídá součtu délek dílčích úseček (Popescu, Mačutek a Altmann 2010). Máme tak

$$L=\sum_{r=1}^{V-1} d(r)=\sum_{r=1}^{V-1} \sqrt{(f(r+1)-f(r))^2+1}.$$

Obdobně určíme délku křivky končící h-indexem. Předpokládejme, že h označuje celočíselnou hodnotu h-indexu. Potom veličina

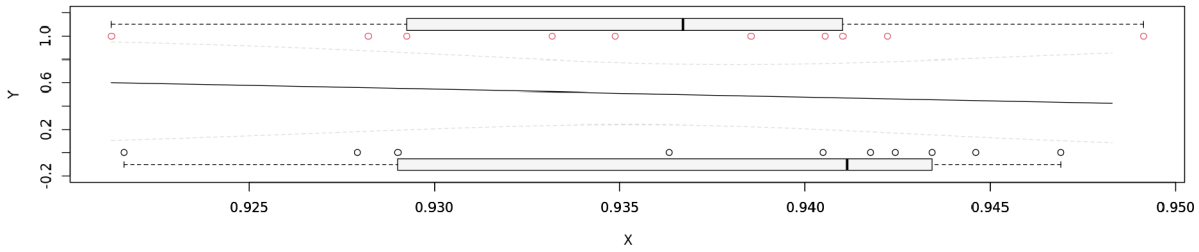
$$H=\sum_{r=1}^{h} d(r)=\sum_{r=1}^{h} \sqrt{(f(r+1)-f(r))^2+1}$$

udává délku křivky končící h-indexem. Poměrnou délku křivky, tedy hledaný r-index, vypočítáme s pomocí vzorce

$$R=\frac{L-H}{L}=1-\frac{H}{L},$$

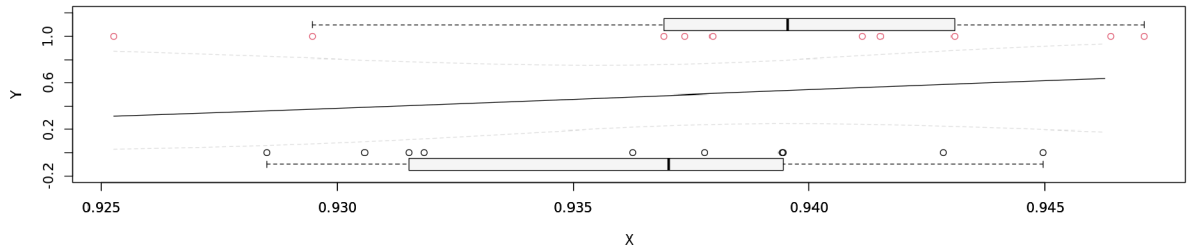
jako poměr výše vypočítané délky křivky H za h-indexem vůči celkové délce křivky. Hodnota r-indexu se tak pohybuje v intervalu mezi čísly nula a jedna (Kubát, Matlach a Čech 2014; Čech, Popescu a Altmann 2014).

V následujících odstavcích se na základě konceptu slovního bohatství, respektive vypočtených hodnot r-indexu, pokusíme ověřit, zda-li je možné jej přímo využít k přiřazení (ověření) autorství. Nejprve s pomocí programu QUITA vypočítáme hodnoty r-indexu pro jednotlivá díla a z těchto hodnot pak logistickou regresí vytvoříme odpovídající rozhodovací model. Při analýze výsledného modelu se, podobně jako v předešlých případech, zaměříme na jeho statistickou významnost s hranicí významnosti 0,01.



***Obrázek 10**: Logistická regrese r-indexu dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 po sobě jdoucích tokenů.*

Na *Obrázku 10* vidíme graf zachycující výsledek logistické regrese r-indexu naměřeného na posloupně vybraném vzorku tokenů. S pomocí získaného modelu můžeme autory k textům správně přiřadit v 55 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí p-hodnoty 0,645286, což znamená, že tento model není funkční, respektive statisticky signifikantní. Zvolený index, tedy r-index při posloupném výběru tokenů, není vhodnou metrikou pro rozlišení autorství námi zkoumaných textů.



***Obrázek 11**: Logistická regrese r-indexu dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 náhodně vybraných tokenů.*

Obrázek 11 přináší výsledný graf logistické regrese r-indexu, tentokrát však měřeného na náhodném vzorku tokenů. S pomocí takto získaného modelu můžeme autory k textům správně přiřadit v 55 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí *p*-hodnoty 0,393202, to znamená, že ani tento model není funkční, respektive statisticky signifikantní. Tento zvolený index tudíž není vhodnou metrikou pro rozlišení autorství námi zkoumaných textů.

3.3.5 Průměrná délka tokenu

Průměrnou délku tokenu (ATL, z anglického Average Token Length), v tabulkách uvedenou jako AVGTOKENLEN, spočítáme pomocí aritmetického průměru, daného vzorcem

$$ATL = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

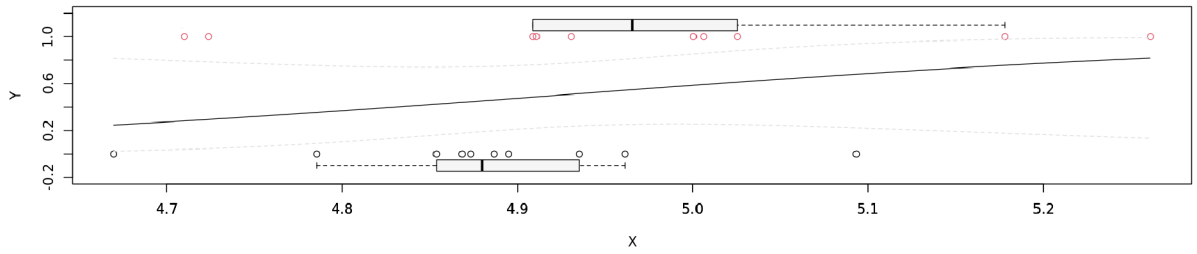
kde proměnné x_i určují délky jednotlivých tokenů v grafémech a symbol N definuje celkový počet tokenů v analyzovaném textu (Kubát, Matlach, Čech 2014).

Délka slov (tokenů) použitých v textu se, v závislosti na jazyce, autorovi a typu literárního díla, řídí obecně různými pravděpodobnostními distribucemi. Nalezením konkrétních parametrů a pravděpodobnostních distribucí ve vztahu k různým jazykům se detailně zabývá mimo jiné Popescu (Popescu et al. 2013). V případě českého jazyka je průměrná délka slova estimovaná na 5,5 grafému (Králík 1983).

Tento textový ukazatel, mimo jiné, souvisí s Menzerath-Altmannovým zákonem, podle kterého je délka jazykového konstruktu úměrná délce jeho konstituentů. Typickým příkladem je souvislost délky slabik s délkou slov: čím více slabik slovo obsahuje, tím jsou použité slabiky v průměru kratší (Čech, Popescu a Altmann 2014). Úměrnost porovnávaných veličin

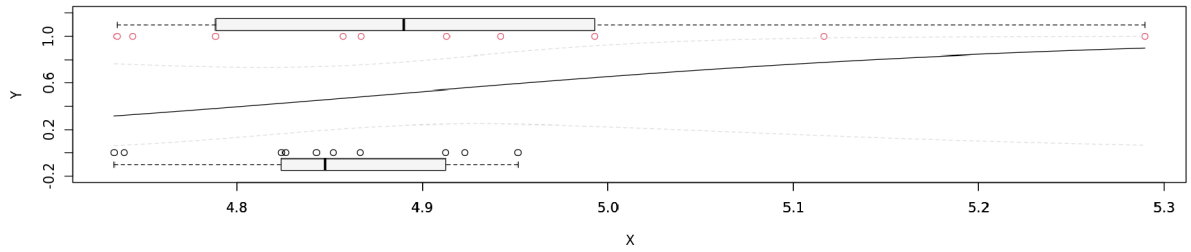
se obecně odvíjí od autorství, typu textu, užitého jazyka, žánru, literárního stylu i samotné formy autorského díla (Čech, Popescu a Altmann 2014). Například v rámci českého jazyka je možné pozorovat závislost průměrné délky slov a frekvenci výskytu jednotlivých grafémů s ohledem na užitý styl: nezanedbatelné rozdíly jsou patrné mezi administrativním, odborným a publicistickým stylem (Králík 1983). V našem případě pracujeme s autorskými texty ve stejném jazyce, žánru i formě. Pro díla jednotlivých autorů by tak měla platit obdobná závislost v kontextu Menzerath-Altmannova zákona.

S tímto předpokladem se pokusíme ověřit, zda-li je možné využít průměrné délky tokenů k přiřazení (ověření) autorství. Nejprve s pomocí programu QUITA vypočítáme průměrné délky tokenů jednotlivých autorských děl a z těchto hodnot pak logistickou regresí vytvoříme rozhodovací model. Při analýze tohoto modelu se, podobně jako v předešlých případech, zaměříme na jeho statistickou významnost s hranicí významnosti 0,01. Při jejím překročení nebudeme model považovat za vypovídající a tedy přínosný při rozhodování o přiřazení autorství.



***Obrázek 12**: Logistická regrese indexu průměrné délky tokenu dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 po sobě jdoucích tokenů.*

Za použití vzorku po sobě jdoucích tokenů jsme naměřili hodnoty indexu zachycujícího průměrnou délku tokenu. S pomocí získaného modelu, zachyceného za *Obrázku 12*, můžeme autory k textům správně přiřadit v 65 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí *p*-hodnoty 0,185382. To znamená, že tento model není funkční, respektive statisticky signifikantní, s ohledem na stanovenou hranici významnosti. Index průměrné délky tokenu měřený na posloupně vybraných tokenech není tudíž vhodnou metrikou pro rozlišení autorství námi analyzovaných textů.



***Obrázek 13**: Logistická regrese indexu průměrné délky tokenu dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 náhodně vybraných tokenů.*

Poslední graf logistické regrese vidíme na *Obrázku 13*. Zobrazuje výsledek pro index průměrné délky tokenu, jehož hodnoty jsme naměřili při náhodném výběru tokenů. *S* pomocí takto získaného modelu můžeme autory k textům správně přiřadit v 60 % případů. Kvalitu tohoto modelu můžeme charakterizovat pomocí *p*-hodnoty 0,170296. Tento model tedy není funkční, respektive statisticky signifikantní, neboť přesahuje stanovenou hranici významnosti. Zvolený index průměrné délky tokenu při náhodném výběru tokenů proto není vhodnou metrikou pro rozlišení autorství textů analyzovaných v této práci.

3.3.6 Výsledky a resumé

V *Tabulce 11* a *Tabulce 12* shrnujeme naměřené výsledky vybraných textových ukazatelů při použití různých metod výběru tokenů.

Při použití metody posloupného výběru tokenů, které jsme vybrali z korpusu dvaceti zkoumaných textů, jejichž délka byla sjednocena na 4500 tokenů (*Tabulka 11*), jsme pomocí lineární regrese zjistili, že při použití ukazatele h-index jsme schopni určit autora textu s pravděpodobností 70 %. Tento výsledek považujeme za signifikantní, neboť vypočítaná *p*-hodnota je nižší, než námi stanovená hraniční hodnota 0,01. Ostatní vybrané textové ukazatele jsme při použití posloupného výběru tokenů na základě jejich vypočítané *p*-hodnoty vyhodnotili jako nesignifikantní.

V *Tabulce 12* jsme za použití metody náhodného výběru tokenů, aplikované na stejný vzorek vstupních dat jako v případě posloupného výběru, zjistili, že signifikantní výsledek byl naměřen pouze v případě textového ukazatele TTR, který při vypočítané *p*-hodnotě 0,0098 umožní stanovit autorství v 65 % případů.

Posloupný výběr tokenů		
Textový ukazatel	**Procentuální účinnost**	***p*-hodnota**
TTR	50 %	0,216063
Entropie	50 %	0,295660
H-index	70 %	0,003505
R-index	55 %	0,645286
AVGTOKENLEN	65 %	0,185382

***Tabulka 11**: Srovnání účinnosti a signifikance vybraných textových ukazatelů získaných lineární regresí při použití metody posloupného výběru tokenů, aplikované na dvacet zkoumaných textů, jejichž délka byla sjednocena na 4500 tokenů. Z vybraných ukazatelů jsme dosáhli signifikantního výsledku pouze pro h-index, při jehož použití jsme schopni rozlišit autorství v 70 % případů.*

Náhodný výběr tokenů		
Textový ukazatel	**Procentuální účinnost**	***p*-hodnota**
TTR	65 %	0,009773
Entropie	70 %	0,060587
H-index	50 %	0,656651
R-index	55 %	0,393202
AVGTOKENLEN	60 %	0,170296

***Tabulka 12**: Srovnání účinnosti a signifikance vybraných textových ukazatelů získaných lineární regresí při použití metody náhodného výběru tokenů, aplikované na stejný vzorek vstupních dat jako v Tabulce 11. Z vybraných ukazatelů jsme dosáhli signifikantního výsledku pouze pro TTR, při jehož použití jsme schopni rozlišit autorství v 65 % případů.*

4 Model Bag-of-Words

Další metodou ověřování autorství, na kterou se zaměříme, je takzvaný Bag-of-Words model, zkráceně BoW. Je jedním ze základních modelů počítačového zpracování textů, který nám umožňuje zcela automaticky, bez našeho zásahu, určit podobnost zkoumaných textů na základě užitých slov. Tento model se často používá při zpracování přirozeného jazyka (NLP). Jeho předností je jeho jednoduchost a s tím spojená relativní výpočetní nenáročnost. Podrobněji o použití modelu Bag-of Words při zpracování dat pojednává například příspěvek z mezinárodní konference v Iráku (Qader, Ameen a Ahmed 2019).

Pokud se rozhodneme vytěžit vlastnosti textu pomocí modelu Bag-of-Words, musíme vzít v potaz, že tento model nijak nezohledňuje pozici slov v textu, jejich vzájemné vztahy v rámci syntaktických a sémantických struktur, ani jejich vzájemnou podobnost, nýbrž spočívá v definování slovníku daného textu. Odtud je odvozen i název modelu. Předmětem zájmu je pouze výskyt slova v textu, nikoliv jeho umístění. Dozvídáme se tak, jaká slova jsou v textu

použita a s jakou frekvencí jsou tato jednotlivá slova zastoupena, tedy počet jejich opakování. Vypočítané hodnoty jsou reprezentovány vektory. Předpokladem potom je, že texty jsou (si) podobné, pokud jsou (si) podobné jejich slovníky. Zkoumáme tedy obsahovou podobnost textů. Zároveň můžeme předpokládat, že takové texty pracují s podobnými tématy.

Aplikací metody vícerozměrného škálování (MDS, popsanou v následující sekci), můžeme určit vzájemnou míru podobnosti zkoumaných textů.

4.1 Metoda vícerozměrného škálování

Jedním z nástrojů, které můžeme využít k analýze zpracovaných dat je metoda vícerozměrného škálování (MDS, z anglického *Multidimensional Scaling)*. S pomocí této metody je možné vizualizovat podobnost mezi jednotlivými prvky množiny různých objektů v obecně vícerozměrném prostoru (Mead 1992).

Jediným požadavkem pro její úspěšné použití je znalost vzdáleností mezi porovnávanými objekty. Musíme tedy pracovat s objekty, jejichž vlastnosti jsou buďto vyjádřené číselně, nebo je možné je na čísla převést. Následně je potřebné určit, jakým způsobem má být vzdálenost mezi jednotlivými objekty vypočítána. Obvykle se pro výpočet vzdálenosti používá Euklidovská vzdálenost (Mead 1992) nebo kosinová nepodobnost (Tan et al. 2019).

Obecně je možné využít metody vícerozměrného škálování při identifikaci témat, která jsou společná zkoumaným textům. K tomuto účelu je možné využít metody singulárního rozkladu (SVD, z anglického Singular Value Decomposition) k identifikaci samotných témat a metody vícerozměrného škálování pro vizuální reprezentaci podobností. Při podrobnější analýze je dokonce možné určit společná témata na úrovni užitého lexika. Patentovaná metoda latentní sémantické analýzy (LSA, z anglického Latent Semantic Analysis), využívaná mimo jiné při zpracování přirozeného jazyka, je ve své podstatě založena na modelu Bag-of-Words. (Kherwa 2017)

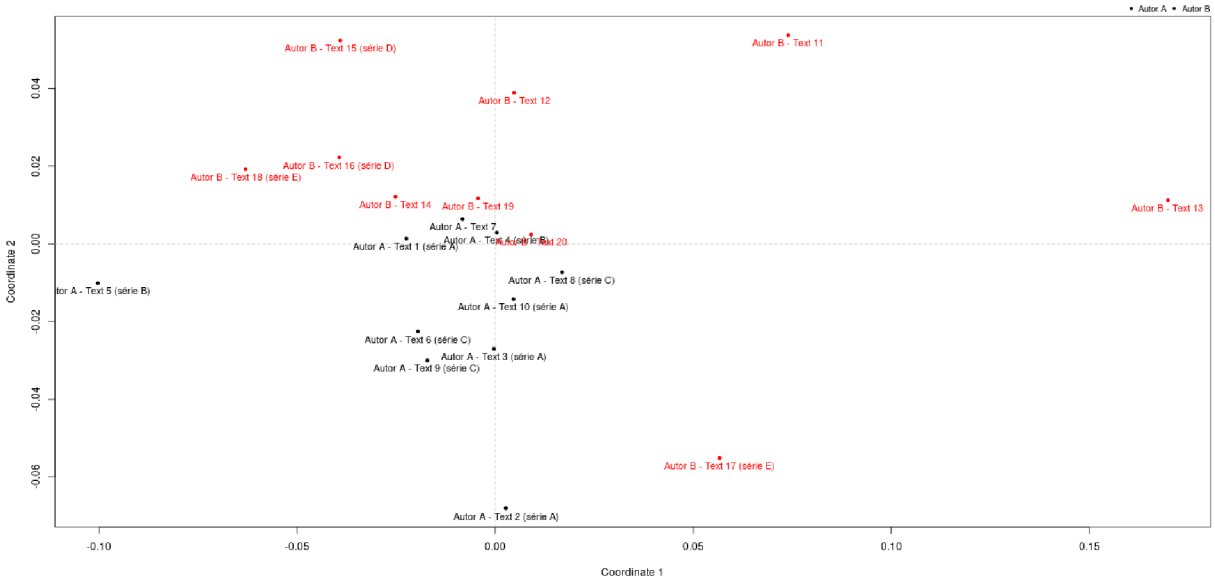
Metody vícerozměrného škálování využijeme pro analýzu (ne)podobnosti jednotlivých textů. Při práci s grafem, který je vykreslený za pomoci této metody, vyhodnocujeme vzájemné rozložení objektů. V našem případě používáme zobrazení do dvourozměrného prostoru, tedy do roviny. Rozmístění jednotlivých bodů, reprezentujících porovnávané objekty, na grafu a jejich vzájemné postavení, respektive vzdálenost, se odvíjí od míry jejich vzájemné podobnosti. Graf tak ve své podstatě tvoří mapu objektů, na které nahlížíme shora. Body nejpodobnějších objektů se budou vzájemně překrývat. Naopak, se vzrůstající nepodobností

se od sebe budou jednotlivé body, které tyto objekty reprezentují, vzdalovat. Kromě vzdálenosti mezi jednotlivými body můžeme dále vyhodnocovat i vzdálenost mezi shluky bodů, do kterých se objekty můžou seskupovat, a rovněž jejich vzájemné postavení.

Velikost níže uvedených obrázků znesnadňuje čitelnost popisků u jednotlivých bodů, ale pro lepší srozumitelnost analýz je dostačující barevná distinkce a dobře patrné rozložení bodů na poli grafu. Podrobně se lze na každý z grafů podívat na konci této práce v sekci Přílohy, kde jsou grafy přiloženy v dostatečném rozlišení (*Obrázek 20* a *Obrázek 21*).

4.2 Příprava a zpracování vstupních dat

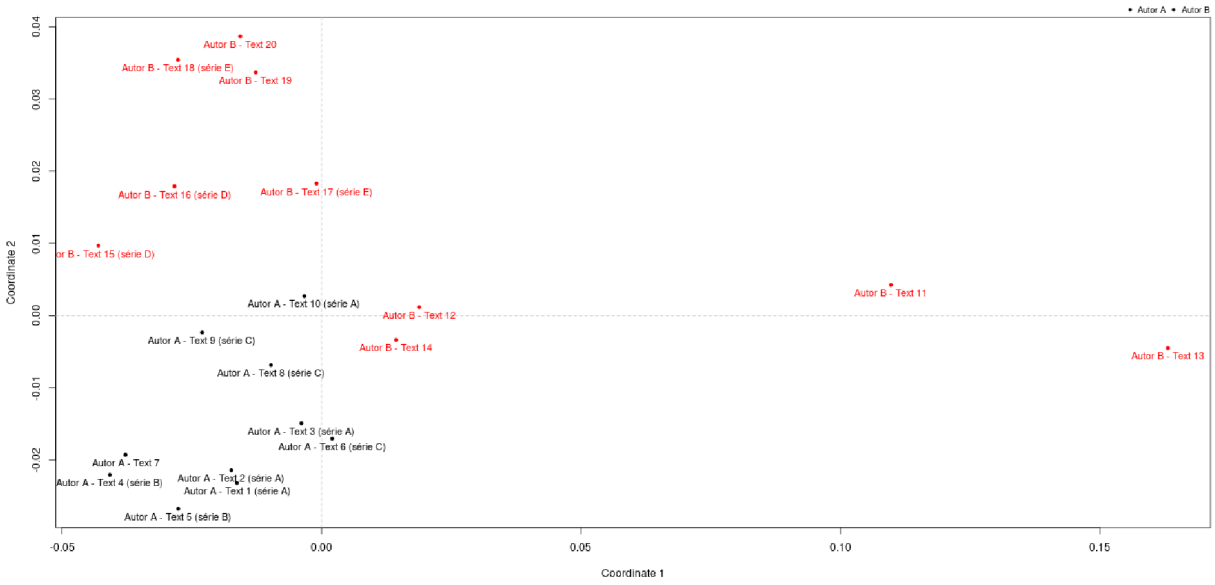
Prvním krokem přípravy a zpracování dat je tokenizace textů. Tu provedeme prostřednictvím výpočetního softwaru QUITA, kde zvolíme pro zpracování *Regex Tokenizer* v jeho výchozím nastavení. V dalším kroku texty ořízneme na námi zvolený jednotný počet 4500 tokenů prostřednictvím nastavení *Tokens Reduce*. Zároveň můžeme zvolit, jestli mají být tokeny vybírány náhodně nebo v tom pořadí, v jakém se vyskytují v textu od jeho začátku. Z těchto vyfiltrovaných výrazů vytvoří QUITA tabulku slov pro každý analyzovaný text. Porovnáním výskytu slov v tabulkách následně určí vzdálenost mezi jednotlivými texty. To provede na základě výpočtu kosinové nepodobnosti. Výsledky získané prostřednictvím modelu Bag-of-Words potom vykreslí do grafu. Pro potřeby další analýzy si zvolíme použití vícerozměrného škálování. V posledním kroku rozhodneme míru podobnosti zpracovaných dat na základě předem stanovených objektivních transparentních kritérií. Ačkoliv je technicky příliš náročné určit proložení přímky grafem a objektivní kritérium homogenity shluků, můžeme grafy proložit pomyslným lineárním separátorem a homogenitu shluků vyhodnotit podle předem stanoveného kritéria. V našem případě jsme zvolili kritérium nadpoloviční většiny.



***Obrázek 14**: Analýza dvaceti zkoumaných textů o velikosti 4500 po sobě jdoucích tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS). Obrázek v plné velikosti je vložen jako příloha této práce (Obrázek 20).*

Na *Obrázku 14* vidíme grafický výstup softwaru QUITA, znázorňující výsledky analýzy 4500 tokenů, které jsme vybrali z dvaceti zkoumaných textů v takovém pořadí, v jakém se vyskytují v původním textu od jeho začátku. Tuto analýzu jsme provedli za použití modelu Bag-of-Words. Pro vykreslení výsledků jsme použili metodu vícerozměrného škálování (MDS). Pro snadnější orientaci v grafu jsou od sebe texty jednotlivých autorů barevně odlišené. Černě vyobrazené texty byly vydány pod jménem *Autora A*, červeně odlišené texty jsou připisované *Autorovi B*. Na první pohled vidíme, že nejhustší koncentrace textů je v místech, kde se protínají horizontální a vertikální osa grafu. Ve větším počtu jsou zde zastoupené černě vykreslené texty *Autora A*. Celkem osm z deseti jeho textů se nachází právě v tomto středovém shluku. Prolínají se s několika červenými body, reprezentujícími texty *Autora B*. V nejtěsnější blízkosti vidíme texty 20, 19 a 14. Vzdálenější, ale stále v dosahu tohoto shluku jsou texty 16 a 18. Ostatní červené body vidíme rozptýlené po celé ploše grafu napravo od středového shluku. Nejvzdálenější jsou texty 11 a 13. Možné vysvětlení, proč se texty 11 a 13 oddělily tak výrazným způsobem od ostatních textů, nalezneme v *Tabulce 2*. Oba texty, se řadí k literárnímu žánru povídka a oba texty mají mnohem menší rozsah, než ostatní texty. Dalším důvodem může být například podobné téma, o kterém tyto texty pojednávají, ale to z výše uvedeného grafu nelze zjistit. Můžeme tedy říct, že zhruba dvě třetiny textů jsou si vzájemně podobné natolik, že tvoří shluk. Zhruba jedna třetina textů je od

ostatních i od sebe navzájem natolik odlišná, že netvoří žádné shluky a jsou rozptýlené po celé ploše pole grafu.



***Obrázek 15**: Analýza dvaceti zkoumaných textů o velikosti 4500 náhodně vybraných tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS). Obrázek v plné velikosti je vložen jako příloha této práce (Obrázek 21).*

Podobně jako v předešlém případě, na *Obrázku 15* vidíme analýzu 4500 tokenů, které jsme vybrali z dvaceti zkoumaných textů. Tentokrát jsme ale použili náhodný způsob výběru tokenů. Ty jsou vybírány z celého původního textu. S ohledem na principy tematické koncentrace textu můžeme očekávat, že výsledky analýzy budou přesnější. Texty jednotlivých autorů jsou barevně rozlišené stejným způsobem jako v předchozím grafu, tedy černě vyobrazené texty náleží *Autorovi A*, červeně vykreslené texty náleží *Autorovi B*. Většina textů obou autorů se nachází v levé části grafu a tvoří jeden shluk. Pouze texty 11 a 13, které jsou připisované *Autorovi B*, se od tohoto shluku oddělily a nachází se na protilehlé straně pole grafu. Když se zaměříme na to, jak jsou uspořádané texty v již zmiňovaném shluku, můžeme vidět, že všechny černé body jsou v jeho spodní části, počínaje textem 10, který leží téměř na průsečíku os grafu, a rozprostírají se v relativně vzájemně těsné blízkosti až po spodní hranici pole. Červené body jsou situované v horní části shluku. Texty číslo 12 a 14 *Autora B* zasahují mezi texty *Autora A*. Důvody, proč se texty 11 a 13 oddělily od shluku ostatních textů tak výrazně, jsou shodné s důvody uvedenými v předchozí analýze. Může to být zapříčiněno jejich literárním žánrem, jejich odlišným rozsahem nebo například tématem, o kterém pojednávají.

4.3 Výsledky a resumé

Provedli jsme analýzu 4500 tokenů, které jsme vybrali dvěma různými způsoby z dvaceti zkoumaných textů, jejichž původ je připisován dvěma různým autorům, které pro účely naší práce označujeme jako *Autora A* a *Autora B*. Tokeny jsme zvolili nejprve v tom pořadí, v jakém se vyskytují v textu od jeho začátku, a následně jsme použili náhodný výběr tokenů z celého textu. Obě varianty jsme podrobili analýze provedené prostřednictvím modelu Bag-of-Words. Výpočet vzdáleností byl proveden za použití kosinové nepodobnosti. Výsledky těchto analýz jsme vizualizovali za pomoci vícerozměrného škálování. Následně jsme se pokusili tyto grafické reprezentace zpracovaných dat interpretovat a popsat.

Z našich popisů grafů lze vyvodit následující závěry. Texty 11 a 13 se výrazným způsobem odlišují od ostatních textů, a to jak v případě posloupného tak náhodného způsobu výběru tokenů. Co je příčinou nedokážeme na základě informací dostupných z těchto dvou grafů přesně určit. V případě posloupného výběru tokenů vnímáme body jako více neuspořádané, v případě náhodného výběru se osmnáct z dvaceti textů, tedy nadpoloviční většina, seskupilo do jednoho shluku. Dále můžeme říct, že výsledky analýzy s náhodným výběrem tokenů vykazují vyšší míru podobnosti mezi jednotlivými texty, než výsledky s posloupným výběrem tokenů.

Ani jeden z grafů nám však neumožňuje rozhodnout, jestli se jedná o texty jedno, dvou nebo více autorů. To může být zapříčiněno tím, že původci zkoumaných textů ve skutečnosti nejsou dva různí lidé. V případě posloupného výběru tokenů by výsledky mohly naznačovat více různých autorů, výsledky náhodného výběru mohou naopak budit dojem, že všechny texty napsala jedna osoba, s výjimkou textu 11 a 13, které mohly být napsány buď jinou osobou, nebo téže osobou, která se záměrně snažila pozměnit svůj obvyklý styl psaní. Druhou možnou příčinou, proč nemůžeme jednoznačně rozhodnout původ zkoumaných textů, je nedostatečnost aplikované metody. Mohlo dojít k tomu, že autory textů jsou dvě různé osoby, ale námi použitá metoda, tedy model Bag-of-Words ve spojení s metodou vícerozměrného škálování a měření vzdáleností kosinovou nepodobností, není dostatečně silná na to, aby rozdíly jejich autorského stylu zachytila. Další možností je vliv délky původního textu na výsledky tohoto experimentu, neboť texty 11 a 13 jsou výrazně kratší než ostatní texty. Protiargumentem tomuto tvrzení by mohla být skutečnost, že text 12 je rovněž výrazně kratší než ostatní texty, dokonce je nejkratším textem z celého korpusu, ale nesdružuje se s texty 11 a 13. Obdobně je to s hypotézou, že se texty 11 a 13 oddělily a shlukly k sobě z důvodu náležitosti ke stejnému literárnímu žánru, neboť jsou to povídky, zatímco ostatní texty

Autora B jsou romány. Text 12 je ovšem rovněž povídka a jak již bylo uvedeno, s texty 11 a 13 se neshlukuje.

5 Hapax legomenon jako ukazatel autorského stylu

Z pohledu korpusové lingvistiky pojmem *hapax legomenon*, dále jen *hapax* nebo *hapaxy* (pl.), označujeme jazykové jednotky, jejichž specifickou vlastností je, že se v korpusu textů vyskytují pouze a právě jen jednou. (Hladká, Novotná a Karlíková 2017) Míra zastoupení těchto jevů v textu potom závisí na jeho délce. V případě, že text bude tvořen počtem slov v řádu jednotek, potom je vysoká pravděpodobnost, že se každé slovo bude vyskytovat v tomto textu pouze jednou a tudíž bude každé takové slovo hapaxem. Zastoupení hapaxů v textu bude tedy 100 %. S rostoucí délkou textu sledujeme postupné snižování míry zastoupení těchto jazykových jednotek. Při velikosti korpusu okolo 5 milionů slov se hladina hapaxů dostane ke svému minimu, zhruba 35 %. Pokud korpus dále rozšiřujeme, míra zastoupení hapaxů začne znovu stoupat, a to až k hladině 55 %. Chování hapaxů v korpusech větších než 120 milionů slov není dosud známo. (Cvrček 2017) Jak vidíme, distribuce hapaxů v závislosti na velikosti korpusu je relativně dobře zdokumentovaný a popsaný jev, jehož hodnoty můžeme predikovat na základě Zipfových zákonů. Podrobně se problematice hapax legomen a jejich vztahu k atribuci autorství věnuje například článek *Hapax legomena jako indikátor autorského stylu a formální znak koheze textu* (Faltýnek, Matlach a Owsianková 2020).

V této kapitole využijeme nízkofrekvenčního lexika hapax legomenon pro srovnání zkoumaných textů (*Tabulka 1* a *Tabulka 2*) a určení jejich podobnosti, respektive příslušnosti k jejich autorům. Podobnost textů budeme analyzovat s pomocí metody hierarchického shlukování a také s dříve popsanou metodou vícerozměrného škálování.

5.1 Hierarchické shlukování

Dalším nástrojem pro analýzu dat je metoda hierarchického shlukování neboli *hierarchical clustering* (Hastie, Tribshirani a Friedman 2009). Tato metoda se využívá v případě, kdy zkoumané objekty vykazují více měřitelných vlastností a my hledáme podobnosti buď mezi objekty samotnými, nebo mezi shluky, do kterých se tyto objekty seskupují. Metoda hierarchického shlukování má několik typů a v našem experimentu budeme využívat typ aglomerativní. To znamená, že výsledná hierarchie podobnosti je tradičně zobrazena jako

strom, tzv. *dendrogram*, a sdružuje k sobě vždy dva nejbližší, tedy nejpodobnější, objekty, případně dva nejbližší shluky. Podobnost je vyhodnocována na základě vypočítaných vzdáleností. Hierarchické shlukování slouží k objektivnímu určení nejen náležitosti objektů do shluků, ale zároveň i určení vzájemné podobnosti zkoumaných objektů. Pro použití hierarchického shlukování musíme nejprve určit metodu, podle které budeme posuzovat vzdálenost mezi shluky, a to specifikací výběru referenta shluku. Nejčastěji používané metriky shlukování zahrnují metodu *Single*, neboli metodu nejbližšího souseda, která pracuje s tím, že referenti jsou dva nejbližší objekty shluku, metodu *Complete*, neboli metodu nejvzdálenějšího souseda, která naopak považuje za referenty dva nejvzdálenější objekty shluku. V metodě *Average* je referentem shluku průměrný objekt. V našich výpočtech užitá *Wardova* metoda (Ward 1963) vycházející z analýzy rozptylu je velmi účinnou metodou, nicméně má tendenci vytvářet poměrně malé shluky.

Vzdálenosti mezi shluky objektivně určují míru podobnosti. Mezi nejčastěji používané metody výpočtu vzdálenosti patří Euklidovská vzdálenost a kosinová nepodobnost, kterou budeme využívat pro hierarchické shlukování. Abychom předešli zkreslení výsledků vlivem různých jednotek vzdálenosti, můžeme využít přeškálování, respektive normalizaci, kdy obvykle využíváme převodu absolutních hodnot na procenta prostřednictvím metody min-max nebo z-skóre. Sledované vlastnosti je rovněž možné ohodnotit, například 1 a 0 pro označení výskytu/absence vlastnosti. Další možností je škálovat míru výskytu.

5.2 Příprava vstupních dat

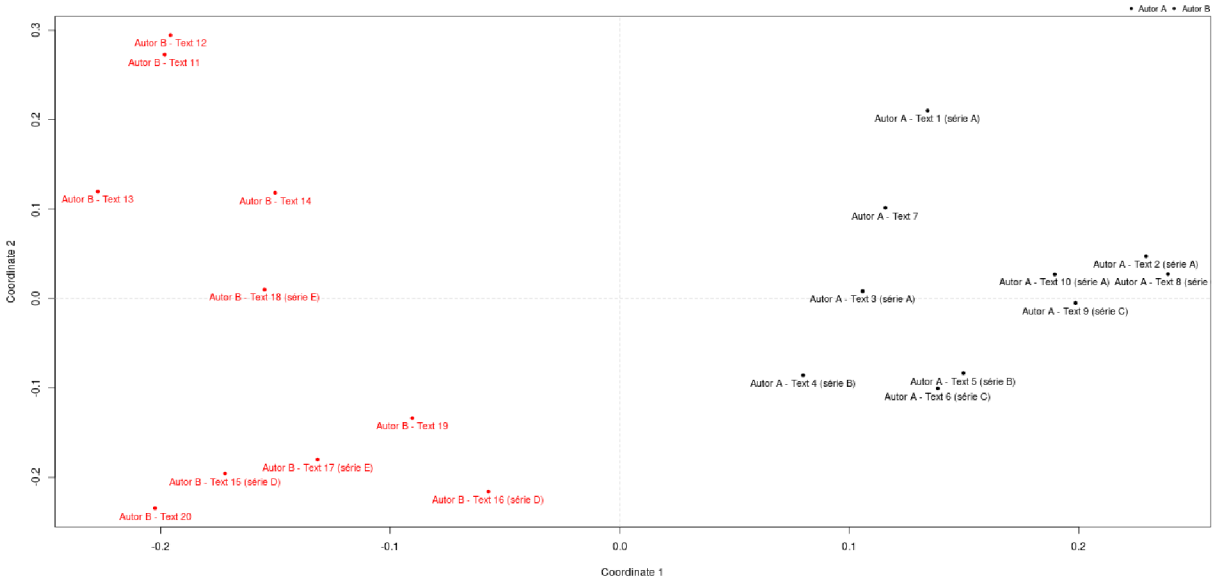
V prvním kroku provedeme tokenizaci textů, stejně jako v případě předchozí metody. I další krok zpracování bude shodný. Texty ořízneme na námi zvolený jednotný počet 4500 tokenů prostřednictvím nastavení *Tokens Reduce*. Rovněž budeme volit, jestli mají být tokeny vybírány náhodně nebo v tom pořadí, v jakém se vyskytují v textu od jeho začátku. Další krok zpracování je klíčový právě pro tuto metodu. Z tokenizovaných textů vybereme pouze výrazy označované jako hapax legomenon, tedy tokeny které se v celém textu vyskytují pouze jednou a tedy mají frekvenci výskytu jedna. To provedeme rovněž prostřednictvím výpočetního softwaru QUITA. Všechny tokeny s frekvencí výskytu vyšší než jedna odfiltrujeme vhodnou volbou *Token Frequency Filter* v nastavení zpracování dat v programu. Tímto způsobem získáme pouze hapaxy. Další zpracování už je opět shodné s předchozí metodou, tedy vykreslení výsledků do grafů prostřednictvím modelu Bag-of-Words.

5.3 Zpracování vstupních dat

Software QUITA nejprve z vybraných hapaxů vytvoří tabulky těchto výrazů pro každý z analyzovaných textů. Tyto budeme dále analyzovat výše popsanou metodou hierarchického shlukování a také s pomocí metody vícerozměrného škálování. V obou případech je rozsah analyzovaných textů omezen na 4500 tokenů, které jsou vybrány dvěma způsoby. Obdobně jako v předešlých kapitolách pracujeme s tokeny získanými v posloupném pořadí, tedy tak, jak se vyskytují v původním textu, a dále s tokeny, které byly z původního textu vybrány náhodně.

Při analýze textů v obou přístupech QUITA vypočítá podobnosti mezi jednotlivými texty s použitím kosinové nepodobnosti. Pro potřeby analýzy využívající hierarchického shlukování využívá Wardovy metody k organizaci jednotlivých textů do stromových grafů, zvaných dendrogramy. Při jejich čtení a popisování budeme postupovat zleva doprava a shora dolů. S metodou vícerozměrného škálování, detailněji diskutovanou v předešlé kapitole, QUITA vytvoří grafy bodů. Na základě shlukování příslušných bodů do skupin, respektive prostorově oddělených struktur, můžeme určit míru podobnosti textů.

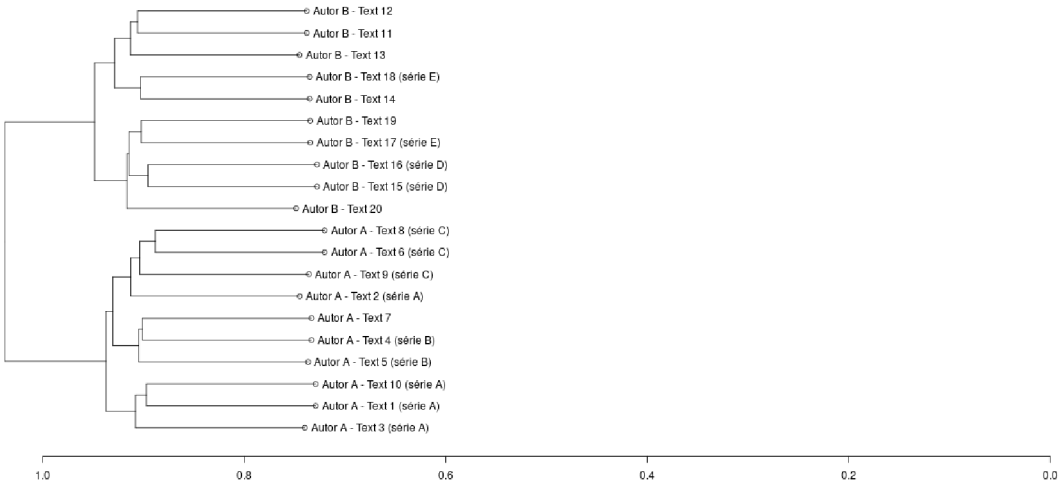
Podrobnosti o jednotlivých použitých textech, tedy jejich zařazení k literárnímu žánru, příslušnost k některé ze sérií a původní rozsah, jsou uvedeny v kapitole s názvem 2.4 Výběr vstupních dat. Níže analyzované grafy jsou rovněž k nahlédnutí ve větším rozlišení v příloze této práce (*Obrázek 22-25*). Předpokládáme, že se v grafech budou texty shlukovat podle jejich náležitosti k autorovi, podle literárního žánru a podle příslušnosti ke knižní sérii.



***Obrázek 16**: Analýza hapax legomen dvaceti zkoumaných textů o velikosti 4500 po sobě jdoucích tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS). Obrázek v plné velikosti je vložen jako příloha této práce (Obrázek 22).*

Obrázek 16 zachycuje grafické znázornění výsledků analýzy hapax legomen, provedenou za použití modelu Bag-of-Words. Na začátku jsme měli 4500 tokenů, které jsme vybrali z dvaceti zkoumaných textů v takovém pořadí, v jakém se vyskytují v původním textu od jeho začátku. Z těchto vybraných tokenů jsme za pomoci softwaru QUITA vyfiltrovali pouze tokeny s frekvencí výskytu 1, tedy hapaxy. Za použití modelu Bag-of-Words jsme provedli výpočet jejich vzdáleností prostřednictvím kosinové nepodobnosti a výsledky vykreslili pomocí metody vícerozměrného škálování. Analyzované texty se shlukují do dvou poměrně jasně oddělených skupin, což nám pomáhá rozlišit vertikální osa grafu. Texty vydané pod jménem *Autora A*, které jsou v grafu rozlišeny černou barvou, se všechny nacházejí v pravé části pole grafu a vzdálenosti mezi nimi jsou natolik malé, že můžeme říct, že tvoří homogenní shluk. Když se podíváme na červeně vykreslené texty, které byly vydané pod jménem *Autora B*, vidíme, že jsou relativně rovnoměrně rozprostřené v podstatě po celé ploše grafu nalevo od vertikální středové osy grafu. Popis rozložení červených bodů by se zřejmě silně odvíjel od subjektivní interpretace těchto dat. Mohli bychom označit dva shluky, jeden nacházející se v levém horním rohu grafu, sestávající z textů 11 a 12, a druhý shluk nacházející se v levé dolní části grafu, do kterého bychom zařadili texty 15, 16, 17, 19, 20. Texty 13, 14 a 18, nacházející se mezi těmito dvěma možnými shluky, bychom mohli označit jako třetí shluk, nebo, s ohledem na zjevně výraznější rozptyl těchto bodů, bychom je mohli

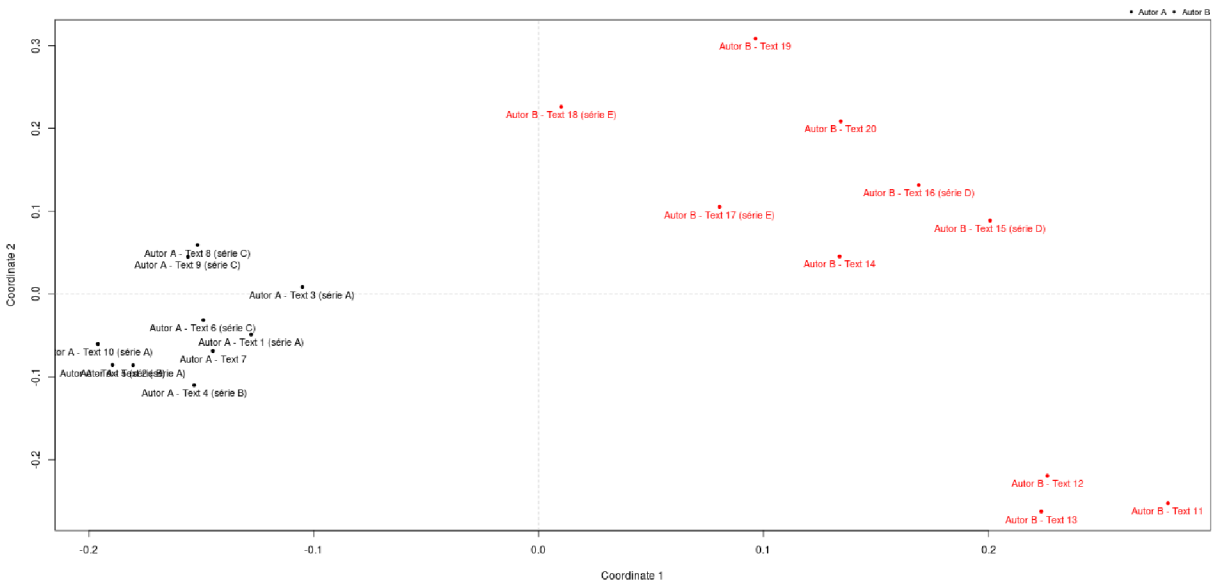
označit za solitérní, bez náležitosti k nějakému shluku. V každém případě nemůžeme o textech *Autora B* jednoznačně říct, že se jedná o jeden homogenní shluk.



***Obrázek 17**: Analýza hapax legomen dvaceti textů o velikosti 4500 po sobě jdoucích tokenů provedená užitím kosinové nepodobnosti. Výsledky jsou vyobrazeny prostřednictvím hierarchického shlukování Wardovou metodou. Obrázek v plné velikosti je vložen jako příloha této práce (Obrázek 23).*

Abychom získali co nejvíce možných informací o zkoumaných textech, provedli jsme znovu analýzu téhož vzorku tokenů metodou Bag-of-Words. Vyfiltrovali jsme tedy hapaxy ze 4500 tokenů posloupně vybraných z dvaceti zkoumaných textů. Pro výpočet vzdáleností jsme použili kosinovou nepodobnost a výsledky jsme zobrazili prostřednictvím hierarchického shlukování Wardovou metodou (*Obrázek 17*). Na první pohled je patrné, že se dendrogram dělí na dvě hlavní větve. Na horní hlavní větvi grafu se sdružují texty náležící výhradně *Autorovi B*, na spodní hlavní větvi grafu se nachází výhradně texty *Autora A*. Můžeme tedy říct, že se texty jednotlivých autorů sdružují do dvou homogenních shluků. Při podrobnějším prozkoumání výsledků vidíme, že texty 11 a 12 jsou si blízké a sdružené na stejné větvi grafu. Z ostatních textů k nim má nejblíže text 13. Při pohledu na parametry původních textů uvedené v *Tabulce 2* vidíme, že pořadí těchto tří textů v grafu koresponduje s jejich vzestupným pořadím hodnot původního rozsahu. Dále je zajímavé, že texty 15 a 16 náleží ke stejné knižní sérii, pro naše účely jsme ji označili jako sérii D, a v našem grafu se nachází na stejné větvi grafu. V případě textů *Autora A* je nejvýraznějším rysem shlukování textů podle sérií na stejných nebo sousedních větvích dendrogramu. Odchylku vidíme pouze v případě textu 2 ze série A, který se oddělil od ostatních textů z této série a přiřadil se na větev

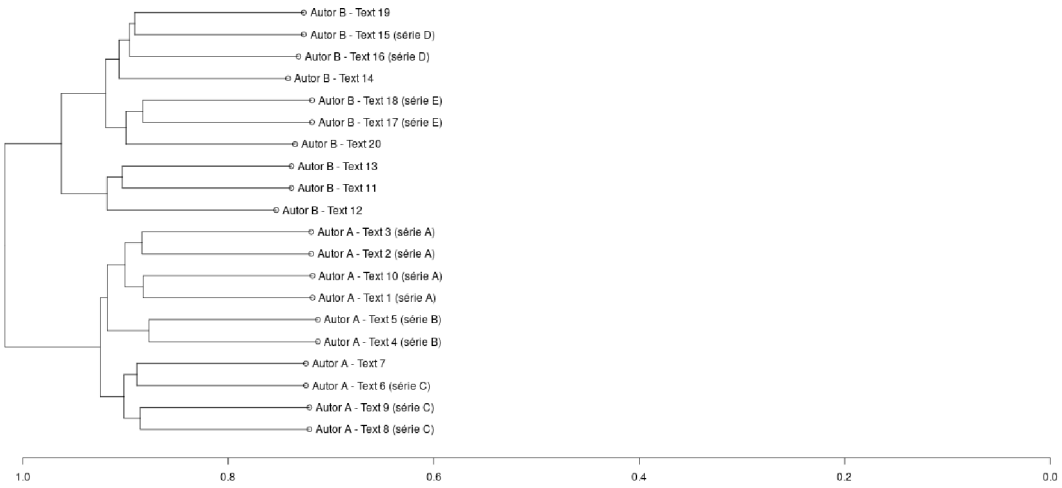
sdružující texty ze série C. Pokud se podíváme na parametry textů, které jsme uvedli v *Tabulce 1*, zjistíme, že všechny tři texty ze série C jsou povídky, zatímco text 2 ze série A je román. Tím tedy můžeme vyloučit možnost, že by tyto texty shlukly podle náležitosti ke stejnému literárnímu žánru. Při porovnání původních délek textů vidíme, že všech deset textů *Autora A* má přibližně stejnou délku, odlišuje se pouze text 9, který je výrazně kratší. Tudíž ani původní rozsah textů nám nijak nenapoví, proč se text 2 podobá textům ze série C. Stejně jako vícerozměrné škálování, ani hierarchické shlukování nám nic nenapoví o tom, jakými tématy se texty zabývají, proto nemůžeme ani v tomto případě vyloučit jejich možnou podobnost, která by vysvětila řazení textů v grafu.



***Obrázek 18**: Analýza hapax legomen dvaceti zkoumaných textů o velikosti 4500 náhodně vybraných tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS). Obrázek v plné velikosti je vložen jako příloha této práce (Obrázek 24).*

Pro srovnání jsme provedli analýzu hapax legomen, vyfiltrovaných ze 4500 tokenů, které byly vybrány náhodným způsobem z původního rozsahu dvaceti zkoumaných textů. Proces zpracování byl v ostatních krocích stejný jako u posloupného výběru. V grafu (*Obrázek 18*) vidíme výsledky vykreslené metodou vícerozměrného škálování. Na první pohled si můžeme všimnout, že se změnilo rozložení červených a černých bodů zrcadlově ve vztahu k vertikální ose grafu. Toto rozdělení je však arbitrární a o textech ani jejich autorství nám nepřináší žádnou novou informaci. Barevné rozlišení zůstává stejné, černě označené texty náleží *Autorovi A*, červeně odlišené texty potom náleží *Autorovi B*. Všech deset textů *Autora A* se nachází ve velmi těsné vzájemné blízkosti a jsou situované v levé části grafu. Nemísí se mezi

ně žádný z textů *Autora B*. Můžeme tedy říct, že texty *Autora A* tvoří homogenní shluk. Při pohledu na červené body nás upoutá jejich rozložení. V pravé spodní části pole grafu vidíme malý shluk textů 11, 12 a 13. O textech 11 a 13 a důvodech jejich vzájemné soudržnosti a odlišnosti od ostatních textů jsme již psali v předchozí části práce v analýzách výsledků zpracování dat metodou Bag-of-Words. Při pohledu na *Tabulku 2* vidíme, že text 12 vykazuje vysokou míru podobnosti s texty 11 a 13 z důvodu zařazení k literárnímu žánru povídka a výrazně menším rozsahem, ve srovnání s ostatními texty tohoto autora. O tématu textů 11, 12 a 13 nemůžeme z našich výsledků nic usoudit, proto ani nemůžeme vyloučit jejich podobnost. V pravém horním kvadrantu vidíme zbylých sedm textů *Autora B*, které se seskupily do dvou téměř rovnoběžných linií. Těchto sedm textů bychom mohli považovat za samostatný shluk.



***Obrázek 19**: Analýza hapax legomen dvaceti textů o velikosti 4500 náhodně vybraných tokenů provedená užitím kosinové nepodobnosti. Výsledky jsou vyobrazeny prostřednictvím hierarchického shlukování Wardovou metodou. Obrázek v plné velikosti je vložen jako příloha této práce (Obrázek 25).*

Poslední analýzou této kapitoly i celé práce je analýza hapax legomen, které byly vyfiltrované ze 4500 tokenů náhodně vybraných z původního rozsahu dvaceti zkoumaných textů. Výpočet vzdáleností byl proveden za užití kosinové nepodobnosti, výsledky byly vykresleny do grafu metodou hierarchického shlukování (*Obrázek 19*). Stejně jako v případě posloupného výběru se nám graf dělí na dvě základní větve, přičemž každá z nich sdružuje pouze texty jednoho z autorů. Shodně se nám texty *Autora B* objevují na horní větvi a texty *Autora A* na spodní větvi grafu. Všechny texty, které náleží k nějaké sérii, se shlukují na stejné nebo sousední větve dendrogramu, tentokrát bez výjimky. Texty 11, 12 a 13 se opět nachází v těsné blízkosti,

v tomto případě seskupené v sestupném pořadí podle původní délky textu, tedy 13, 11, 12. Zaujímají spodní větev shluku textů *Autora B*.

5.4 Výsledky a resumé

V této kapitole jsme se zaměřili na nízkofrekvenční lexikum, označované jako hapax legomenon, a jeho využití při zkoumání autorství textů. Provedli jsme čtyři analýzy, jejichž vstupní data byla vybrána z dvaceti různých textů, které jsou připisované dvěma různým autorům. Texty jsme nejdříve za pomoci softwaru QUITA tokenizovali, následně jsme vybrali 4500 tokenů. V prvním případě jsme volili tokeny v jejich původním pořadí od začátku textu, v druhém případě jsme použili náhodný výběr. Z nich jsme následně vyfiltrovali pouze hapaxy a jejich vzdálenosti jsme měřili pomocí kosinové nepodobnosti. Výsledky jsme vizualizovali dvěma různými způsoby, a to prostřednictvím modelu vícerozměrného škálování a modelu hierarchického shlukování. Tímto způsobem jsme získali čtyři grafy, které jsme podrobili analýze.

Z těchto grafů jsme získali mnoho zajímavých informací. Prvořadě vidíme, že bez ohledu na způsob výběru tokenů se texty dělí do dvou základních větví, přičemž každá z nich nese výhradně texty jednoho z autorů. Tyto větve se dále dělí a vytváří celou strukturu vztahů zkoumaných dat. Nejvíce nápadná je zřejmě tendence textů 11 a 12, potažmo 13, držet se vzájemně v těsné blízkosti, ať už ve shluku (*Obrázek 16* a *Obrázek 18*) nebo větvi dendrogramu (*Obrázek 17* a *Obrázek 19*). Další významný jev je shlukování textů podle náležitosti ke knižním sériím, ve kterých byly vydány. Tato skutečnost je lépe patrná z grafů hierarchického shlukování. V případě posloupného výběru tokenů se k sobě neshlukly pouze texty ze série E a oddělil se text 2 ze série A (*Obrázek 17*). Při výběru tokenů náhodným způsobem už k žádným odloučením nedošlo a všechny texty se shlukly podle náležitosti k sériím, ve kterých byly vydány (*Obrázek 19*).

6 Výsledky analýz

V první ze tří praktických částí jsme se zaměřili na použití textových ukazatelů při určování autorství. V prvním kroku tohoto experimentu jsme prostřednictvím faktorové analýzy vybrali pět textových ukazatelů, u nichž jsme dále zjišťovali, s jakou měrou pravděpodobnosti můžeme právě na jejich základě přiřadit autorství daného textu. Tuto vlastnost zkoumaných indexů jsme vyhodnocovali za použití modelu logistické regrese. Pro porovnání jsme provedli výběr vstupních dat (tokenů) dvojím způsobem, a to posloupně a náhodně, tudíž máme dvě sady výsledků, které jsme uvedli v *Tabulce 11* a *Tabulce 12*. Z nich vyplývá, že při posloupném výběru tokenů je h-index jediným textovým ukazatelem, na základě kterého dokážeme rozhodnout o autorství našich zkoumaných textů, a to s určitostí 70 %. Při náhodném výběru tokenů jsme schopni přiřadit autorství pouze na základě indexu TTR, který nám dává jistotu 65 %, že jsme autora přiřadili správně.

Na vykreslení výsledků analýz modelu Bag-of-Words a hapax legomen jsme aplikovali metodu vícerozměrného škálování. Dostali jsme grafy, které nám umožnily vyvodit obecné závěry, avšak bez možnosti podrobnějšího zkoumání vztahů mezi jednotlivými objekty. To znamená, že na první pohled jsme mohli rozhodnout, které vzorky jsou si blízké na základě toho, že spolu tvoří shluky. Dále můžeme určit počet těchto shluků, jejich vzájemné postavení a vzdálenost mezi nimi. Čím větší je podobnost jednotlivých vzorků, tím těžší je odlišit od sebe jednotlivé body v grafu, které tyto vzorky reprezentují, neboť jejich vzájemné vzdálenosti se budou se vzrůstající podobností zmenšovat, a to až natolik, že se můžou vzájemně překrývat. Totéž platí i pro již zmiňované shluky. Na základě našich výsledků, které vidíme na *Obrázku 14* a *Obrázku 15*, můžeme vyhodnotit podobnost jednotlivých analyzovaných textů mezi sebou, ale nedokážeme rozhodnout, jestli jsou tyto texty dílem jednoho, dvou nebo více autorů.

Při zpracování analýz hapax legomen jsme pro porovnání výsledků vytvořili i sérii grafů prostřednictvím metody hierarchického shlukování. Podle rozložení textů na větvích dendrogramů můžeme porovnat jejich blízkost, respektive podobnost. Můžeme identifikovat shluky, do kterých se texty sdružují, a jejich vzdálenost, obdobně jako při čtení výsledků z grafů vícerozměrného škálování. Dendrogramy nám poskytují především podrobnější pohled na vztahy mezi jednotlivými vzorky. Můžeme tak zkoumat, jestli se texty shlukují například podle náležitosti ke knižním sériím, ve kterých byly vydány. Tento faktor podobnosti totiž není z grafů vícerozměrného škálování zřetelně patrný.

Na základě námi provedených analýz, ve kterých jsme se zaměřili na nízkofrekvenční lexikum hapax legomenon jako ukazatel autorského stylu, jsme došli k závěru, že texty vydané pod jménem *Autora A* mají tendenci shlukovat se do vzájemně těsnějších blízkostí a sdružovat se podle náležitosti k jednotlivým knižním sériím. Texty *Autora B* byly více rozptýlené na ploše grafu a sdružování textů do sérií, které rovněž můžeme pozorovat, nepůsobí tak významně zřejmě kvůli jejich menšímu početnímu zastoupení ve srovnání s počtem knižních sérií *Autora A*. Co ovšem můžeme sledovat u textů *Autora B* jako poměrně výrazný rys, je tendence řazení textů podle jejich původní délky při použití metody hierarchického shlukování. Texty 11, 12 a 13, které jsou výrazně kratší, než ostatní texty tohoto autora, se shlukují na téže větvi dendrogramu, ať už jsou analyzované tokeny vybrány z textů v jejich původním pořadí nebo náhodně (*Obrázek 17* a *Obrázek 19)*. Rozdíl je pouze v jejich pořadí, kdy při náhodném výběru tokenů jsou tyto texty seřazeny sestupně (*Obrázek 19)*, a při zachování původního sledu tokenů jsou v dendrogramu vykresleny na jeho první větvi ve vzestupném pořadí od nejkratšího po nejdelší z nich (*Obrázek 17)*. I v grafech vícerozměrného škálování vidíme, že se body reprezentující texty 11, 12 a 13 nachází ve vzájemné blízkosti, v případě náhodného výběru tokenů tvoří dokonce samostatný shluk (*Obrázek 18*).

7 Diskuze

Po tom, co jsme viděli průběh a výsledky jednotlivých analýz, provedených na základě vybraných kvantitativních metod používaných při určování autorství, se nabízí celá řada otázek. Zejména nás zajímá, jestli dokážeme zodpovědět stěžejní otázku, která nás provází celou naší prací, a která zní: Dokážeme od sebe rozlišit texty dvou různých autorů pomocí některé z aplikovaných metod? V návaznosti na to nás zajímá, jestli jsou tyto metody dostatečně silné samy o sobě nebo jestli je musíme kombinovat, abychom dosáhli jistých výsledků. Rovněž se můžeme ptát, který z testovaných přístupů nám dává nejlepší, tzn. statisticky signifikantní, výsledky. A v neposlední řadě je potřeba se ptát, čím mohou být výsledky ovlivněny.

Obecně lze při hledání odpovědí na otázky týkající se validity výsledků kvantitativně lingvistického experimentu uvažovat nad několika problematickými momenty, ke kterým mohlo dojít. Prvním z nich je možnost, že použitý vzorek dat není reprezentativní, tj. bylo použito malé množství vstupních dat nebo byly pro další zpracování vybrány příliš krátké úseky textů, což by mělo za následek, že se zkoumané zákonitosti textů a výskyty jazykových jevů neprojevily v analyzovaném vzorku v plné míře nebo dokonce vůbec. Další možností je, že byla zvolena úplně nevhodná vstupní data, ať už ve vztahu k metodě, která na ně bude aplikována, nebo vzájemně vůči sobě. Příkladem takto nevhodně sestaveného korpusu by mohlo být kombinování textů různého žánru. Opačný problém může nastat, pokud máme adekvátně sestavený soubor vstupních dat, ale pro jejich další zpracování zvolíme nevhodné metody. Pokud máme správně zvolená jak vstupní data, tak metodu jejich zpracování, a přesto nedostáváme odpovídající výsledky, může to mimo jiné znamenat, že zvolené metody nejsou dostatečně silné.

Ve vztahu k našemu experimentu je v první řadě potřeba říci, že výsledky měření, kde je použit náhodný výběr tokenů, jsou pouze ilustrativní a slouží k ukázce toho, jakým způsobem modely s daty zachází a jaké tendence můžeme sledovat ve výsledcích, které nám poskytují. Pro přesnější výsledky by bylo nutné provést výpočty s náhodným výběrem tokenů v nějakém statisticky významném počtu opakování a z těchto dílčích výsledků následně za použití vhodných statistických metod vyvodit finální výsledek.

Primárním cílem naší práce bylo ověření funkčnosti použitých metod a vyhodnocení jejich účinnosti ve vztahu ke zkoumaným datům. Dále byl stanoven předpoklad, že se texty budou

shlukovat podle náležitosti k autorovi, podle kritéria literárního žánru a podle náležitosti ke knižní sérii. Otázkou tedy je, zda dokážeme na základě získaných výsledků rozhodnout se statistickou významností, jestli se jedná o texty jednoho, dvou nebo více autorů a jestli můžeme sledovat a pojmenovat tendence, podle kterých se k sobě zkoumané texty shlukují.

Při pohledu na výsledky prvního experimentu, zkoumajícího informační přínos vybraných textových ukazatelů se musíme ptát, čím může být způsobeno, že nám vyšly takové výsledky. Za předpokladu, že jsme vybrali vhodné indexy, použili jsme je na vhodná data a pracovali jsme s nimi správným způsobem, můžou naše výsledky znamenat, že jsou si texty vzájemně natolik podobné, respektive že naměřené indexy, s výjimkou h-indexu pro posloupný výběr tokenů a indexu TTR pro náhodný výběr tokenů, vykazují natolik podobné hodnoty, že nedokážeme s určitostí rozhodnout o autorství zkoumaných textů.

Model Bag-of-Words, druhý z testovaných přístupů k určování autorství, nebere v potaz syntaktickou strukturu textu, nevybírá tokeny podle slovních druhů, frekvence, pořadí ve větě či valence, ani není kontextově vázaný. Přesto jeho prostřednictvím dokážeme rozhodnout o tom, které texty jsou si vzájemně podobné. Z výsledků našeho experimentu však nedokážeme jednoznačně rozhodnout o počtu autorů analyzovaných textů. Vliv na to neměl ani způsob výběru tokenů z původních textů. Jak jsme již zmínili, výsledek pro náhodný výběr tokenů je pouze ilustrativní. Zůstává proto otázkou, a otevřenou možností pro pokračování v tomto experimentu, jestli by provedení adekvátního počtu opakování a statistické vyhodnocení takto získaných dat vedlo ke stejnému závěru, nebo by přineslo jiný výsledek.

V neposlední řadě nás zajímalo, jestli může být užitečné provést vizualizaci výsledků modelu Bag-of-Words aplikovaného na hapax legomenon dvěma různými způsoby, konkrétně prostřednictvím vícerozměrného škálování a hierarchického shlukování. Zaměřili jsme se na to, jestli nám tyto dvojí výsledky přinesly rozdílné informace o zpracovávaných datech, které by nám pomohly při rozhodování o původu textu. Obecně lze říci, že z výsledků vyobrazených prostřednictvím vícerozměrného škálování je lépe patrné vzájemné postavení a vzdálenost jednotlivých textů, zatímco přínosem hierarchického shlukování je struktura vzájemné podobnosti zkoumaných dat. Proto jsme došli k závěru, že nahlížet na téže výsledky tímto dvojím způsobem není nadbytečné, jak se mohlo zpočátku zdát, ale že se tyto metody vzájemně doplňují a poskytují nám podrobnější náhled na zpracovávaná data a umožňují jejich komplexnější analýzu. Stejně jako v předchozím experimentu, i zde byly výsledky analýz provedených na základě náhodného výběru tokenů použity pouze jako ilustrativní.

Domníváme se však, že tendence, které sledujeme v těchto výsledcích, by byly zachovány při libovolném počtu opakování takových měření.

8 Závěr

V úvodu této práce jsme zvolili tři kvantitativní metody, konkrétně textové ukazatele, model Bag-of-Words a nízkofrekvenční lexikum hapax legomenon. Tyto metody jsme aplikovali na sady dat, které jsme vyfiltrovali z dvacet vybraných česky psaných beletristických textů vydaných pod jmény dvou různých autorů. Zaměřili jsme se na využití těchto metod z pohledu problematiky ověřování autorství a pokusili jsme se porovnat jejich přínos a statisticky vyhodnotit jejich účinnost.

Představili jsme teoretické pozadí prováděných experimentů. Zadefinovali jsme termíny a východiska, se kterými se čtenář v naší práci setkává. Rovněž jsme představili nástroje, které jsme použili v našich experimentech (2.3 Použitý software). Můžeme konstatovat, že jsme v průběhu provádění všech experimentů nenarazili na žádné problémy, které by nám po technické stránce znemožnily jejich provedení. Dále jsme čtenáře seznámili s tím, podle jakého klíče jsme zvolili vstupní data (2.4 Výběr vstupních dat) a jak z nich vybíráme vzorky, které následně zpracováváme a analyzujeme (2.5 Zpracování a příprava vstupních dat).

V první ze tří praktických částí jsme prostřednictvím faktorové analýzy vybrali pět textových ukazatelů. Na předem definovaném vzorku textů jsme provedli měření těchto indexů, k čemuž jsme použili software QUITA. Naměřené hodnoty jsme zanesli do dvou tabulek (*Tabulka 9* a *Tabulka 10*), rozdělených podle způsobu výběru tokenů. Následně jsme hodnoty každého z těchto textových ukazatelů vložili do aplikace pro zpracování logistické regrese. Jejím výstupem jsou grafické vizualizace a procentuální vyhodnocení úspěšnosti tohoto modelu, uvedené v podkapitolách věnovaných jednotlivým textovým ukazatelům. Sdružené výsledky tohoto experimentu jsme shrnuli do *Tabulky 11* a *Tabulky 12,* kterými uzavíráme tuto část práce.

V další sekci jsme stručně představili model Bag-of-Words, který jsme vybrali jako druhý z testovaných přístupů k ověřování autorství. Spolu s ním jsme v této části krátce uvedli i metodu vícerozměrného škálování, kterou jsme použili pro vizualizaci výsledků provedeného experimentu. Vyhodnocení těchto výsledků je do jisté míry subjektivní, neboť se odvíjí od prostorového vnímání a interpretace grafů, nicméně jsme se snažili o určitou objektivizaci zavedením kritéria nadpoloviční většiny při rozhodování o signifikantnosti naměřených výsledků.

V poslední prakticky zaměřené části jsme znovu pracovali s modelem Bag-of-Words, ale tentokrát jsme ho aplikovali pouze na nízkofrekvenční lexikum hapax legomenon vyskytující se v analyzovaném vzorku dat. To, co jsme považovali za zajímavé, bylo porovnání informačního přínosu dvou různých metod pro vizualizaci výsledků našich měření. Vedle již použitého vícerozměrného škálování jsme aplikovali metodu hierarchického shlukování a zaměřili jsme se na to, jestli nám tento model může přinést o zkoumaných datech jiné nebo nové informace, které by byly užitečné při rozhodování o původu analyzovaných textů.

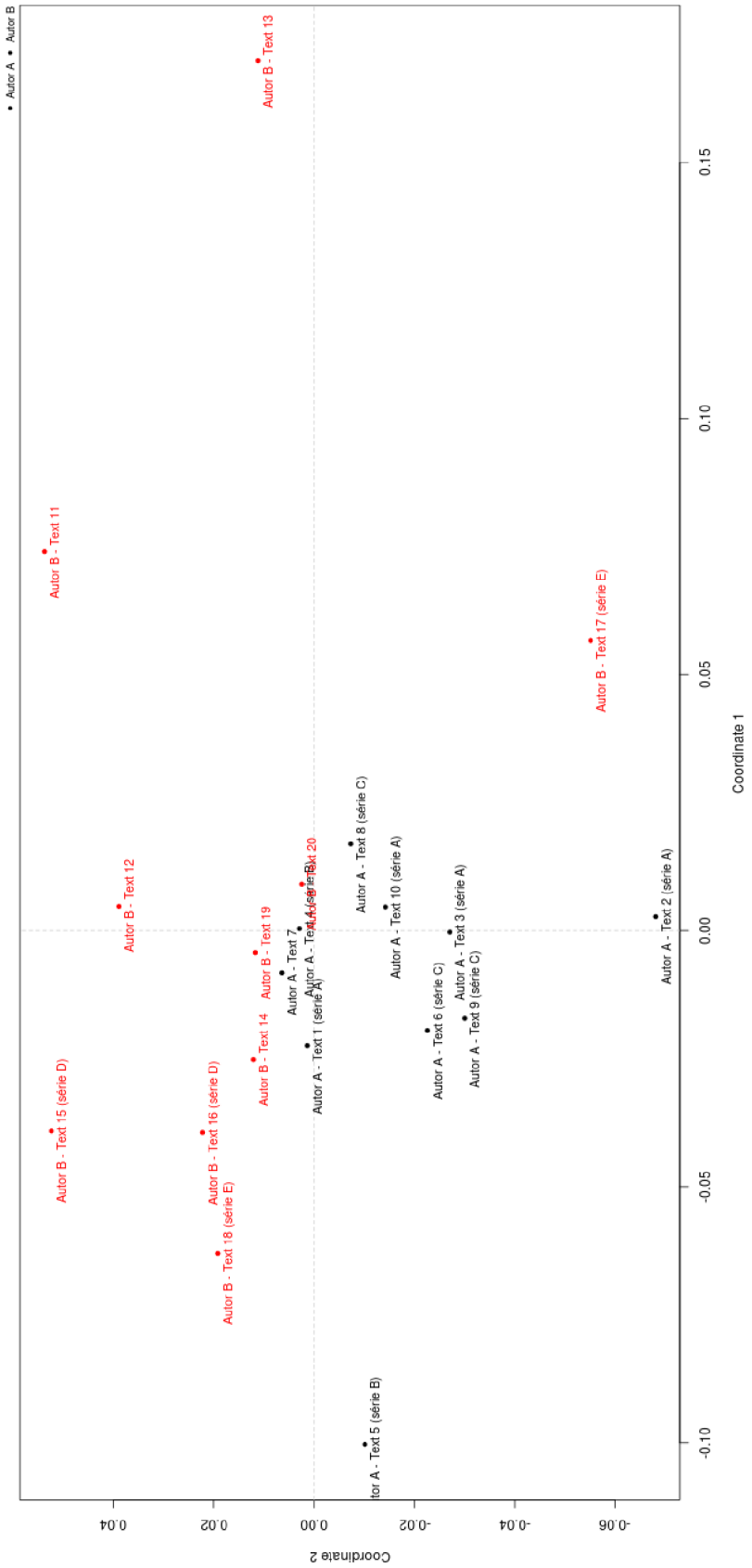
Každá praktická část byla zakončena shrnutím provedeného experimentu a prezentací jeho výsledků. Pro přehlednost jsme však považovali za vhodné zařadit za poslední experiment kapitolu, kde jsme prezentovali souhrnné výsledky celé naší práce.

V kapitole 7 Diskuze jsme posléze vyhodnotili výsledky našich experimentů. Z pohledu určování autorství jsme porovnali mezi sebou informační přínos jednotlivých aplikovaných metod, navrhli jsme možná vylepšení provedených experimentů, položili jsme si otázky, které vyplývají z našich výsledků a domníváme se, že by mohly posloužit jako námět pro další výzkum, respektive rozšíření a vylepšení našich experimentů.

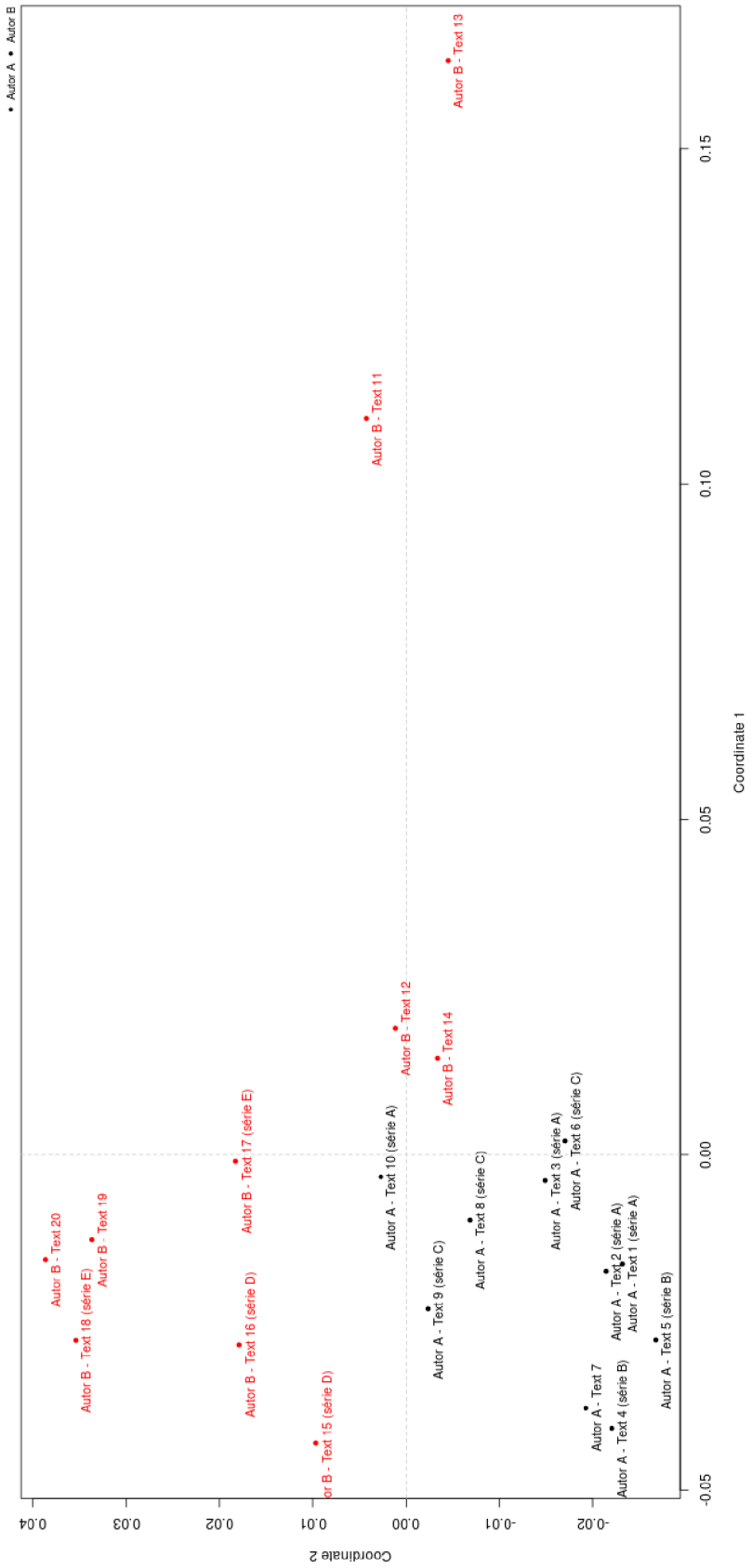
Na základě výše uvedených závěrů si dovolujeme tvrdit, že cíle práce, které jsme v jejím úvodu stanovili, byly naplněny.

9 Přílohy

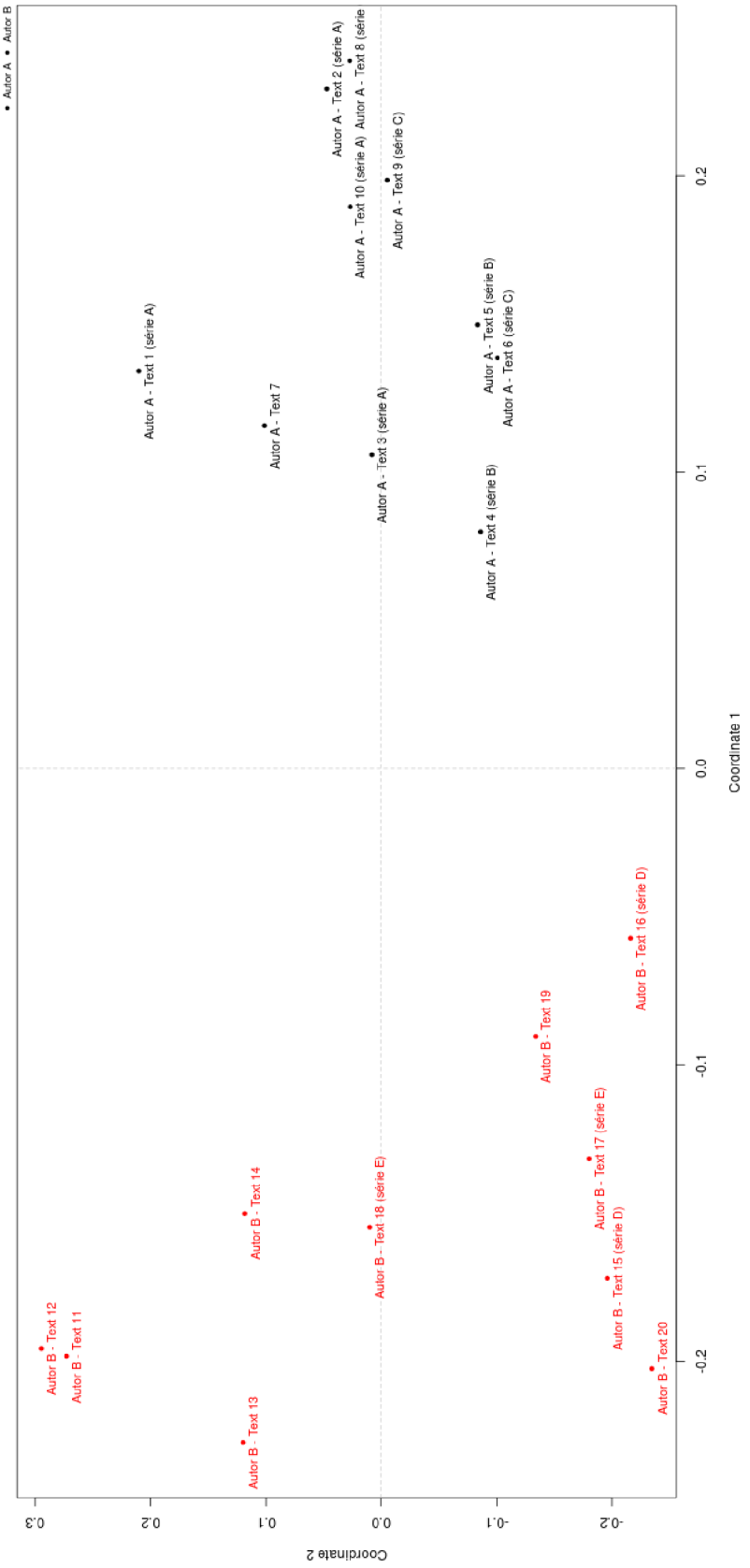
V této části práce pro přehlednost uvádíme některé z diskutovaných grafů ve vyšším rozlišení.



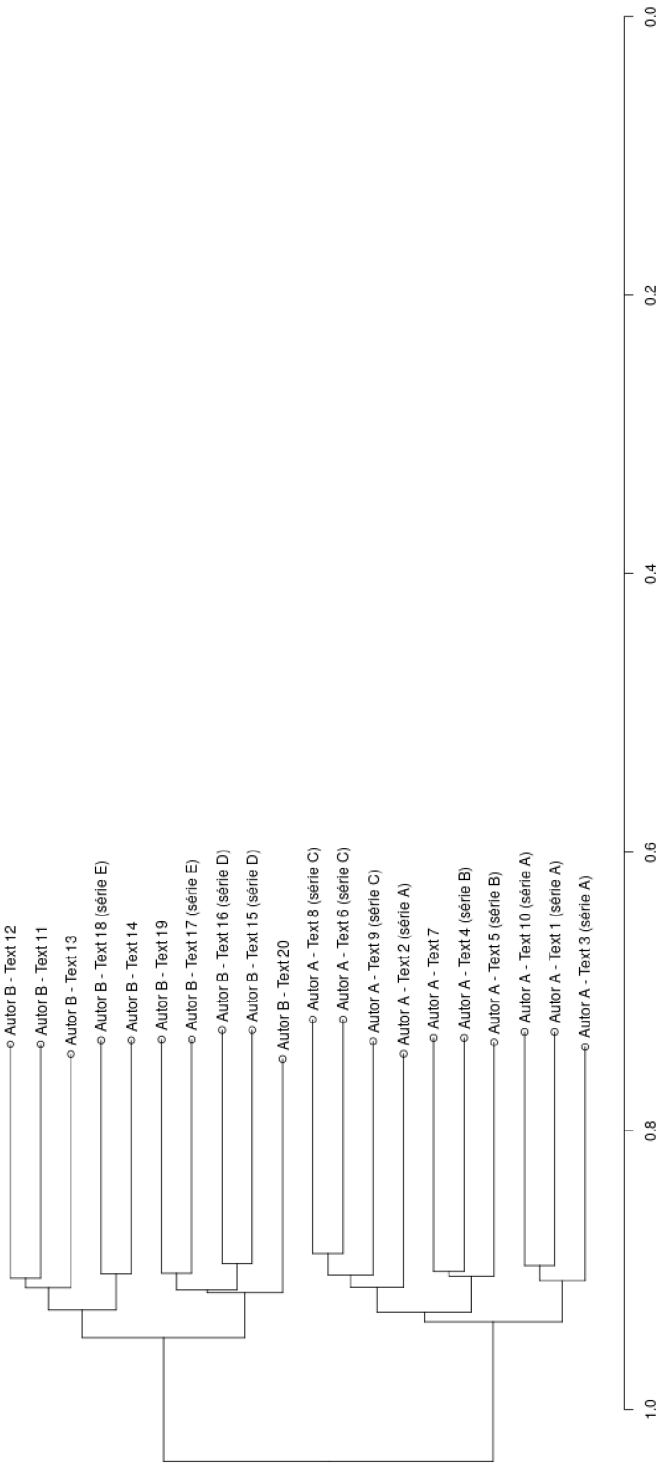
***Obrázek 20**: Analýza dvaceti zkoumaných textů o velikosti 4500 po sobě jdoucích tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*



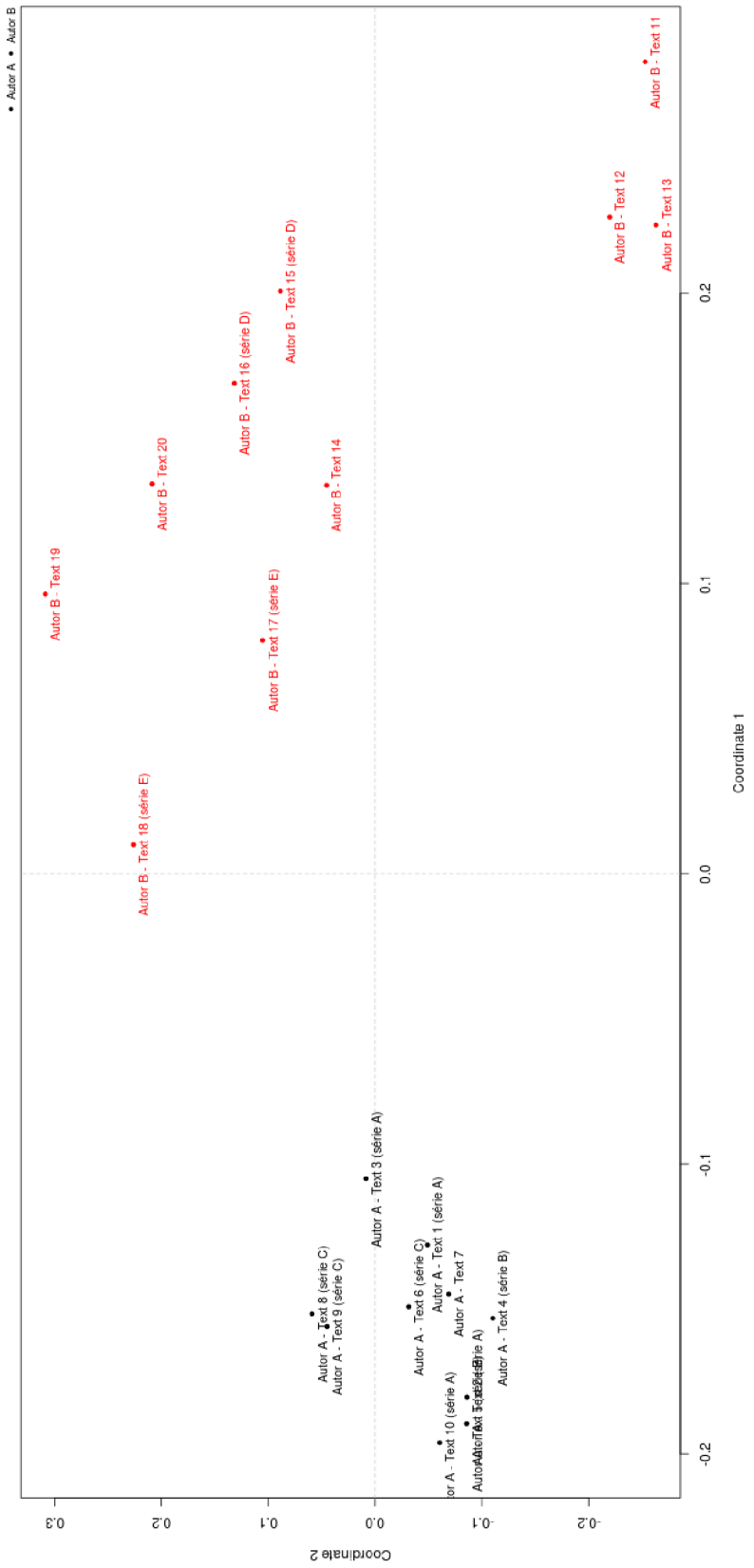
***Obrázek 21**: Analýza dvaceti zkoumaných textů o velikosti 4500 náhodně vybraných tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*



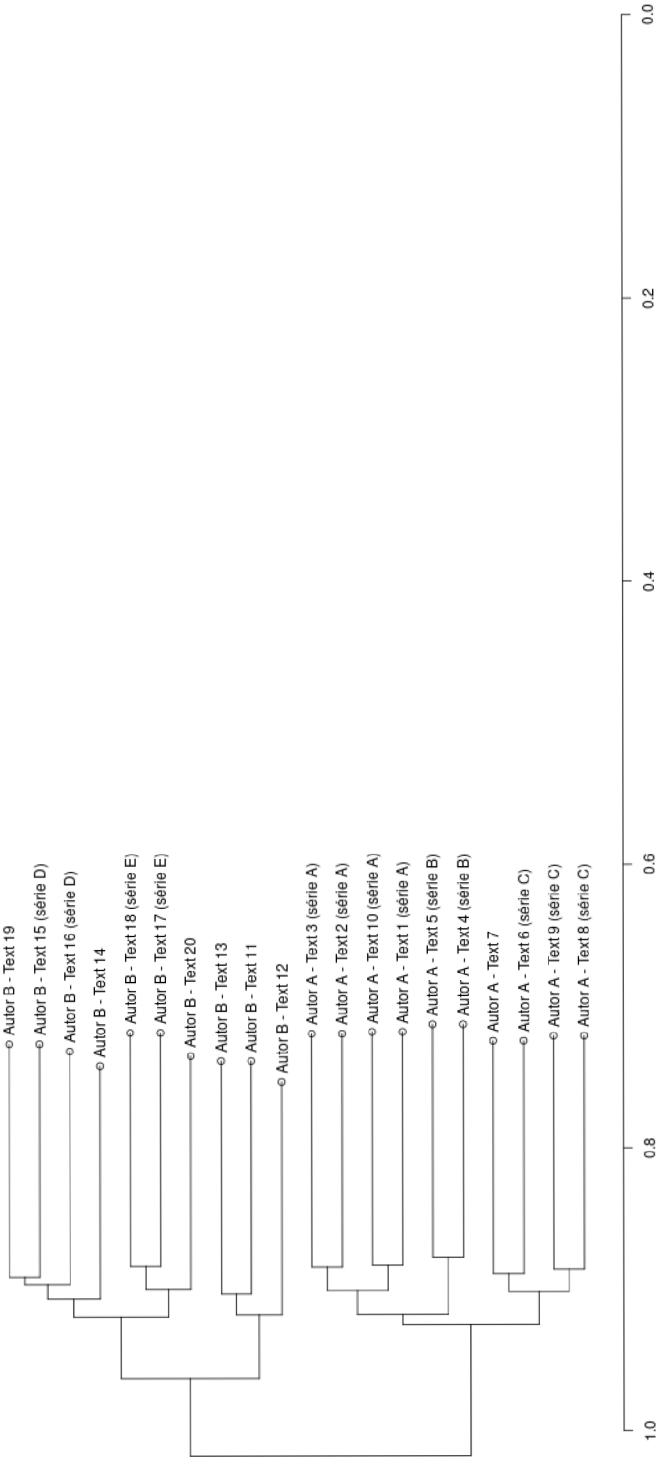
***Obrázek 22**: Analýza hapax legomen dvaceti zkoumaných textů o velikosti 4500 po sobě jdoucích tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*



***Obrázek 23**: Analýza hapax legomen dvaceti textů o velikosti 4500 po sobě jdoucích tokenů provedená užitím kosinové nepodobnosti. Výsledky jsou vyobrazeny prostřednictvím hierarchického shlukování Wardovou metodou.*



***Obrázek 24**: Analýza hapax legomen dvaceti zkoumaných textů o velikosti 4500 náhodně vybraných tokenů, provedená za použití modelu Bag-of-Words. Výsledky jsou vyobrazeny prostřednictvím vícerozměrného škálování (MDS).*



***Obrázek 25**: Analýza hapax legomen dvaceti textů o velikosti 4500 náhodně vybraných tokenů provedená užitím kosinové nepodobnosti. Výsledky jsou vyobrazeny prostřednictvím hierarchického shlukování Wardovou metodou.*

10 Bibliografie

Brémaud, Pierre. 1988. *An Introduction to Probabilistic Modeling.* Springer: New York.

Constantin, Cristinel. 2015. Using the Logistic Regression Model in Supporting Decisions of Establishing Marketing Strategies. *Bulletin of the Transilvania University of Brasov. Series V: Economic Sciences,* 8(2). 43-50.

Cover, Thomas M. a Joy A. Thomas. 2006. *Elements of Information Theory.* John Wiley & Sons, Inc.: Hoboken.

Cramer, Jan Salomon. 2002. The Origins of Logistic Regression. *Tinbergen Institute Discussion Paper,* 2002-119/4.

Crook, Jonathan N., David B. Edelman a Lyn C. Thomas. 2007. Recent Developments in Consumer Credit Risk Assessment. *European Journal of Operational Research*, 183(3). 1447-1465. DOI: 10.1016/j.ejor.2006.09.100

Cvrček, Václav. 2017. Hapax. V: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny.*

Cvrček, Václav. 2017a. Token. V: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny.*

Čech, Radek, Ioan-Iovitz Popescu a Gabriel Altmann. 2014. *Metody kvantitativní analýzy (nejen) básnických textů*. Univerzita Palackého v Olomouci: Olomouc.

Čech, Radek. 2016. *Tematická koncentrace textu v češtině.* Ústav formální a aplikované lingvistiky: Praha.

Čech, Radek a Miroslav Kubát. 2017. Slovní bohatství textu. V: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny.*

Doležel, Lubomír. 1963. Předběžný odhad entropie a redundance psané češtiny. *Slovo a slovesnost*, 24(3). 165-175.

Dreiseitl, Stephan a Lucila Ohno-Machado. 2002. Logistic Regression and Artificial Neural Network Classification Models: a Methodology Review. *Journal of Biomedical Informatics*, 35(5-6). 352-359. DOI: 10.1016/S1532-0464(03)00034-0

Faltýnek, Dan, Vladimír Matlach a Hana Owsianková. 2020. *Hapax legomena jako indikátor autorského stylu a formální znak koheze textu.* DOI: 10.13140/RG.2.2.16509.79847

Fano, Ugo. 1957. Description of States in Quantum Mechanics by Density Matrix and Operator Techniques. *Reviews of Modern Physics,* 29(1). 74-93. DOI: 10.1103/RevModPhys.29.74

Hastie, Trevor, Robert Tribshirani a Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer: New York.

Hirsch, Jorge E. 2005. An Index to Quantify an Individual's Scientific Research Output. *Proceedings of the National Academy of Sciences*, 102(46). 16569-16572. DOI: 10.1073/pnas.0507655102

Hladká, Zdeňka, Renata Novotná a Helena Karlíková. 2017. Hapax legomenon. V: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny.*

Holmes, David I. 1994. Authorship Attribution. *Computers and the Humanities*, 28. 87-106. DOI: 10.1007/BF01830689

Hosmer, David W. a Stanley Lemeshow. 2000. *Applied Logistic Regression*. John Wiley & Sons, Inc.: New York.

Juola, Patrick. 2008. Authorship Attribution. *Foundation and Trends in Information Retrieval*, 1(3). 233-334. DOI: 10.1561/1500000005

Jurka, Michal a Dan Faltýnek. 2017. Forenzní ligvistika. V: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny.*

Kherwa, Pooja a Poonam Bansal. 2017. Latent Semantic Analysis: An Approach to Understand Semantic of Text. *International Conference on Current Trends in Computer, Electrical, Electronics and Communication.* 870-874. DOI: 10.1109/CTCEEC.2017.8455018

Králík, Jan. 1983. Statistika českých grafémů s využitím moderní výpočetní techniky. *Slovo a slovesnost,* 44(4). 295-304.

Krámský, Jiří. 1959. Teorie sdělné promluvy. *Slovo a slovesnost*, 20(1). 55-66.

Kubát, Miroslav, Vladimír Matlach a Radek Čech. 2014. *QUITA – Quantitative Index Text Analyser*. RAM-Verlag: Lüdenscheid.

Malinowski, Edmund R. 2002. *Factor Analysis in Chemistry*. John Wiley & Sons, Inc.: New York.

Manning, Christopher D. a Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing.* The MIT Press: Cambridge.

Mead, Andrew W. 1992. Review of the Development of Multidimensional Scaling Methods. *The Statistician*, 41(1). 27-39. DOI: 10.2307/2348634

Owsianková, Hana, Dan Faltýnek a Ondřej Kučera. 2018. Genetic Analysis of Cabbages and Related Cultivated Plants Using the Bag-of-Words Model. *Linguistic Frontiers*, 1(2). 122-132. DOI:10.2478/lf-2018-0011

Petkevič, Vladimír. 2017. Tokenizace. V: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny.*

Popescu, Ioan-Iovitz a Gabriel Altmann. 2006. Some Aspects of Word Frequencies. *Glottometrics*, 13. 23-46.

Popescu, Ioan-Iovitz. 2007. Text Ranking by the Weight of Highly Frequent Words. V: Peter Grzybek a Reinhard Köhler (eds.), *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday,* 555-566. De Gruyter Mouton: Berlín, Boston. DOI: 10.1515/9783110894219.555

Popescu, Ioan-Iovitz. 2009. *Word Frequency Studies*. De Gruyter Mouton: Berlín, New York. DOI: 10.1515/9783110218534

Popescu, Ioan-Iovitz, Ján Mačutek a Gabriel Altmann. 2010. Word Forms, Style and Typology. *Glottotheory*, 3(1). 89-96. DOI: 10.1515/glot-2010-0006

Popescu, Ioan-Iovitz, Sven Naumann, Emmerich Kelih, Andrij Rovenchak, Haruko Sanada, Anja Overbeck, Reginald Smith, Radek Čech, Panchanan Mohanty, Andrew Wilson a Gabriel Altmann. 2013. Word Length: Aspects and Languages. V: Reinhard Köhler a Gabriel Altmann (eds.), *Issues in Quantitative Linguistics,* 3. 224-281. RAM-Verlag: Lüdenscheid.

Qader, Wisam A., Musa M. Ameen a Bilal I. Ahmed. 2019. An Overview of Bag of Words; Importance, Implementation, Applications, and Challanges. *International Engineering Conference on Developments in Civil and Computer Engineering Applications.* 200-204. DOI: 10.1109/IEC47844.2019.8950616

Quaia, Emilio a Federica Vernuccio. 2022. The H Index Myth: A Form of Fanaticism or a Simple Misconception? *Tomography*, 8(3). 1241-1243. DOI: 10.3390/tomography8030102

Sedlačíková, Blanka. 2012. *Historie matematické lingvistiky.* Akademické nakladatelství CERM v Brně: Brno.

Shannon, Claude E. 1948. The Mathematical Theory of Communication. *Bell System Technical Journal*, 27. 379-423.

Spearman, Charles. 1904. General Intelligence Objectively Determined and Measured. *American Journal of Psychology*, 15(2). 201-293. DOI:10.2307/1412107

Svobodová, Marie. 1997. Forenzní lingvistika: obsah a možnosti. *Slovo a Slovesnost*, 58. 124-129.

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne a Vipin Kumar. 2019. *Introduction to Data Mining*. Pearson: Velká Británie.

Uhlířová, Ludmila. 2017. Kvantitativní lingvistika. V: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny.*

Ward, Joe H. 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58. 236-244.

Zhao, Ying, Justin Zobel a Philip Vines. 2006. Using Relative Entropy for Authorship Attribution. *Proceedings of the 3rd Asia Retrieval Symposium AIRS*, 92-105. DOI: 10.1007/11880592_8