Filozofická fakulta Univerzity Palackého v Olomouci

Katedra obecné lingvistiky

Užití disperze nízkofrekventovaného lexika k určení autorství textu

magisterská diplomová práce

Autor: Bc. Dan Valihrach

Vedoucí práce: Doc. Mgr. Dan Faltýnek, Ph.D.

Olomouc

2024

Prohlášení

Prohlašuji, že jsem magisterskou diplomovou práci „Název diplomové práce“ vypracoval/a samostatně a uvedl/a jsem veškerou použitou literaturu a veškeré použité zdroje.

V Olomouci dne ……. Podpis …….

Abstrakt

Název práce: Užití disperze nízkofrekventovaného lexika k určení autorství textu

Autor práce: Bc. Dan Valihrach

Vedoucí práce: Doc. Mgr. Dan Faltýnek, Ph.D.

Počet stran a znaků: 79 stran, 111 690 znaků

Počet příloh: 4

Abstrakt: Tato diplomová práce se zabývá využitím vlastností lexikální disperze slov v textu pro potřeby určování autorství textu. Zkoumá vlastnosti disperze slov v rámci korpusu textů beletrie a snaží se na základě způsobu disperze klasifikovat lexikum do tříd. K popisu disperze je zde využit čistě kvantitativní přístup. Atribuční potenciál různých skupin lexika určených pomocí jejich disperze a frekvence je analyzován pomocí výpočtu matic nepodobnosti a aplikací metod hierarchického shlukování a vícerozměrného škálování. V rámci této práce analyzujeme texty korpusu beletrie, které zpracujeme použitím vlastních programovacích skriptů. Cílem této práce je tedy nalézt lexikum, pomocí kterého lze nejlépe identifikovat autorství textu, a zároveň analyzovat možnosti využití lexikální disperze jako stylometrické vlastnosti. V této práci je věnována zvýšená pozornost lexiku s nízkou frekvencí výskytu, na rozdíl od v podobných výzkumech běžně používaného vysoce frekventovaného lexika.

Klíčová slova: disperze, nízkofrekventované lexikum, určování autorství, stylometrie, kvantitativní lingvistika, superhapax

Abstract

Title: Using low-frequency lexical dispersion to determine text authorship

Author: Bc. Dan Valihrach

Supervisor: Doc. Mgr. Dan Faltýnek, Ph.D.

Number of pages and characters: 79 pages, 111 690 characters

Number of appendices: 4

Abstract: This thesis focuses on using features of lexical dispersion of words in texts for the purposes of authorship attribution. It analyses the features of words dispersion in a corpora of fiction books, and attempts to classify words into classes based on their lexical dispersion. Dispersion is analyses using purely quantitative approach. To analyse the potential of various lexical groups for authorship attribution we use dissimilarity matrix computation and application of hierarchical clustering method and multidimensional scaling method. In this thesis we analyse corpora of fiction books, which we process using our own programming scripts. The goal of this thesis is to identify words with the highest potential to distinguish authorship and also to research viability of a lexical dispersion as a stylometric feature. This thesis pays extra attention to the low frequency words, as opposed to high frequency words, which are often used in this field of research.

Keywords: dispersion, low frequency lexis, authorship attribution, stylometry, quantitative linguistics, superhapax

Obsah

1. Úvod

Tato práce se zabývá analýzou disperze nízkofrekventovaného lexika v textech za účelem určení autorství textů. Většina výzkumů v oblasti určování autorství se soustředí především na vysoce frekventované lexikum tvořené funkčními slovy, zatímco lexiku s nízkofrekventovanou frekvencí nebyla věnována dostatečná pozornost. V rámci této práce se pokusíme navázat na výzkum Dana Faltýnka z Univerzity Palackého v Olomouci, který se zabývá studiem vlastností nízkofrekventovaného lexika a zejména průběžně se v textu opakujících semihapaxů, označovaných jako tzv. superhapaxy. Chování nízkofrekventovaného lexika v autorském textu bylo statisticky evaluováno Vladimírem Matlachem. V rámci našeho výzkumu se budeme zabývat vztahem rozmístění výskytů slov v textu, tedy disperze slov, a jejich frekvence výskytů. Pomocí analýzy těchto textových vlastností se pokusíme identifikovat lexikum s nejvýhodnějšími vlastnostmi pro odlišení autora textu od jiných autorů.

Tato diplomová práce je složena ze dvou hlavních části. První z těchto dvou částí je část teoretická. V jejím rámci se budeme zabývat určováním autorství jako oblastí výzkumu. Zmíníme zde ranné počátky tohoto odvětví, následný vývoj a postupy užívané v současnosti. Zvláštní pozornost budeme věnovat metodám využívajícím nízkofrekventované lexikum. Analyzujeme společné vlastnosti některých metod a uvedeme, s jakými typy problémů určování autorství se lze v praxi setkat. Po teoretické části navážeme částí praktickou. Zde představíme vlastní přístup k problému určování autorství užitím analýzy disperze lexika. Nejprve představíme naši hypotézu a postup při výzkumu. Navážeme představením naší metody. Poté se budeme zabývat klasifikací lexika do námi definovaných tříd popisujících způsob, jakým je toto lexikum rozprostřeno v rámci textu. Praktickou část zakončíme testováním vlastností naší metody na korpusu textů. Závěrem práce shrneme získané poznatky a nastíníme případné směřování dalšího bádání.

2. Teoretická část

2.1. Počátky stylometrie

Počátky stylometrie jsou některými badateli datovány do poloviny 19. století. Bývají spojovány s britským matematikem Augustem De Morganem, který se v roce 1851 v dopise adresovanému Reverendu W. Healdovi zabývá možností využití průměrné délky slov pro určení sporného autorství epištol přisuzovaných svatému Pavlovi. První experimenty s určováním autorství za pomoci stylometrických metod přišly až později ke konci 19. století. Spojují se s americkým fyzikem Thomasem Corwinem Mendenhallem. V roce 1887 vydal článek s názvem „The Characteristic Curve of Composition“, kde navrhuje využití analýzy křivek relativních frekvencí slov různé délky pro rozlišení autorství textu. Později v roce 1901 publikoval článek „A Mechanical Solution to a Literary Problem“, kde aplikuje svůj přístup na skutečné případy rozporovaného autorství. Těmito případy byly texty přisuzované Williamu Shakespearovi, Francisu Baconovi a Christopheru Marlowovi. Podobnou metodu použil nezávisle na Mendenhallovi i William Benjamin Smith, který v roce 1888 v článku „Curves of Pauline and Pseudo-Pauline Style“, posuzuje autorství epištol svatého Pavla, přičemž obdobně jako Mendenhall porovnává tvař křivek různých textových vlastností. (Plecháč, 2021)

Ve století dvacátém došlo k významnému vývoji v rámci stylometrie a určování autorství. Významnou zásluhu na tomto vývoji měl objev George Kingsleyho Zipfa, který v roce 1932 formuluje tzv. Zipfův zákon. Tento zákon stanoví, že ve všech přirozených jazycích lze najít u slov stejnou distribuci jejich frekvence a pořadí ve frekvenčním seznamu. Tento objev měl významný dopad jak na odvětví zabývající se studiem určování autorství, tak i na oblast celé kvantitativní lingvistiky. (Plecháč, 2021) (Zipf, 1949)

Jako další se o rozvoj stylometrie ve dvacátém století zasadil George Udny Yule. Z počátku prosazoval pro porovnání autorských stylů využití délek vět měřených v počtech slov. V rámci tohoto výzkumu se zabýval zkoumáním vlastností křivek distribucí délek vět posuzovaných autorů. Později ve svém díle „The Statistical Study of Literary Vocabulary“ uvedl novou metriku určenou k měření bohatství slovníku daného autora, kterou označujeme jako „Yuleho K“. (Plecháč, 2021) (Yule, 1944)

2.2. Současné metody určování autorství

2.2.1 Burrowsova delta a její variace

Řada metod aplikovaných na problémy určování autorství v současnosti využívá tzv. míry Delta navržené Johnem Burrowsem v roce 2001. Míra Delta měří vzdálenost mezi jedním textem a skupinou textů. Skupina textů zde reprezentuje styl daného období či žánru textů. Tento typ analýzy využívá ne pouze jednu vlastnost, ale více vlastností textu. Jde tedy o multivariační analýzu. Nejčastěji bývají jako vlastnosti k analýze zvoleny frekvence 50 – 100 nejčastěji opakovaných slov v textech. Burrowsova Delta vykazuje dobré výsledky na různých žánrech textů a lze ji aplikovat i na jiné jazyky než pouze na angličtinu. Burrowsova Delta se stala běžně užívanou a akceptovanou metodou. Burrowsovu Deltu můžeme chápat jako míru vzdálenosti reprezentací textů coby vektorů ve vícerozměrném prostoru, kde každé ze zvolených slov či vlastností odpovídá jednomu z rozměrů v tomto prostoru. Prvním krokem je vytvoření matice textů a zvolených slov, která obsahuje frekvence těchto slov. Následně jsou frekvence standardizovány pomocí z-skóre tak, aby nabývaly pouze hodnot mezi 0 a 1. Burrowsova Delta pro dva texty je pak definována jako $\Delta_B = \sum_{i=1}^{n} |z_i(D_1) - z_i(D_2)|$, tedy jako suma standardizovaných rozdílů frekvencí všech zvolených slov mezi dvěma texty. Vedle Burrowsovy delty byly jinými badateli navrženy alternativní míry využívající odlišné způsoby výpočtu. Řadí se mezi ně kupříkladu lineární Delta, kvadratická Delta, anebo rotovaná Delta, které navrhl Shlomo Argamon, ale existuje i řada dalších návrhů na vylepšení Burrowsovy Delty od jiných badatelů. (Evert, a další, 2017)

2.2.2 Nízkofrekventované lexikum

Většina metod využívajících multivariační analýzu, mezi které se řadí výše zmíněná Burrowsova Delta, obvykle vybírá k analýze pouze malý počet vysoce frekventovaných funkčních slov. Tato volba je opodstatněná skutečností, že funkční lexikum nebývá silně ovlivněno kontextem a bývá rovněž vysoce odolné vůči jazykovým změnám. Některé výzkumy však poukazují na to, že lexikum s nízkou frekvencí může též obsahovat informaci užitečnou k určení autorství textu.

Na možnost využití autorsky specifického lexika k určení autorství textu vedle běžně užívaných vysoce frekventovaných funkčních slov poukazuje ve své práci například George Mikros. Experimentálně porovnává úspěšnost různých výběrů vlastností, kdy vůči sobě srovnává výběry frekventovaného funkčního lexika, autorsky specifického lexika a souboru textových proměnných běžně užívaných k určovaní autorství. Autorsky specifické lexikum definuje jako soubor n slov, který obsahuje soubor k nejvíce distinktivních obsahových slov pro každý text v korpusu. Výsledky experimentu ukazují, že užití autorsky specifického lexika může přinést signifikantní zlepšení přesnosti metody při určování autorství. (Mikros, 2009)

Nízkofrekventovaným lexikem se také zabývá Dan Faltýnek a Vladimír Matlach a další výzkumníci z Univerzity Palackého v Olomouci, na jejichž výzkum se v rámci této diplomové práce pokoušíme navázat. Pro určení autorství navrhují využití tzv. hapax legomen a semihapax legomen. Hapax legomenon je termínem označujícím lexikum vyskytující se v textu právě jedenkrát. Za semihapax můžeme označit lexikum vyskytující se v textu s frekvencí velmi blízkou jednomu výskytu. Autoři si všímají, že některé nízkofrekventované lexikum je typické pro určitého autora a může tedy být použito k určení autorství. Pro rozlišení autorství na základě hapax legomen postupují následovně: Prvním krokem je náhodný výběr určitého počtu hapax legomen z posuzovaných textů. Následně jsou tyto hapaxy převedeny do tzv. bag-of-words reprezentace, kdy jsou texty reprezentovány vektory nul a jedniček značícími přítomnost či nepřítomnost slov v globálním slovníku. Na základě těchto vektorů je vypočítána matice kosinových nepodobností textů, načež jsou nepodobnosti textů vizualizovány pomocí metody vícerozměrného škálování. Výsledky tohoto postupu dokazují, že nízkofrekventované lexikum lze použít k určování autorství textů. (Faltýnek & Matlach, Hapax Remains: authorial features of textual cohesion in authorship attribution, 2020)

Vedle absolutních hapax legomen se autoři zabývají i lokálními hapaxy. Jedná se o takové lexikum, které můžeme považovat za hapax legomenon v případě, kdy za celý text považujeme pouze úsek o určité délce pocházející z původního textu. Z pohledu celého textu se pak nejedná o absolutní hapax legomenon, nýbrž o průběžně se opakující nízkofrekventované lexikum. Autoři pro toto lexikum zavádí termín „superhapax“. Všímají si, že právě superhapax se podílí na spoluvytváření autorského stylému. Charakterizují tak autora jako jedinečného uživatele jazykového systému.

Superhapax může zahrnovat jak nízkofrekventovaná funkční slova, tak i z větší části slova obsahová. To poukazuje na tendenci autora opakovaně vybírat obdobné způsoby strukturace textu a na tendenci opakovaně s nízkou frekvencí tematizovat určité skutečnosti. Jedinečná kombinace těchto nízkofrekventovaných jazykových prostředků se ukazuje jako distinktivní pro daného autora, čímž umožňuje odlišení jeho textů při srovnání s texty jiných autorů. (Faltýnek & Kučera, Parasyntax jako struktura nízko frekventovaných částí textu Hapax legomenon prostředkem textové koheze, 2022)

2.3. Metody užívané k určení autorství textu

K řešení různých typů problémů určování autorství byly historicky využívány různé metody a postupy. Přestože se navzájem značně liší, můžeme je rozdělit do tří různých skupin na základě způsobu, jakým pracují s textovými vlastnostmi. První a nejstarší skupinou metod jsou metody, které analyzují jednu invariantu. Určují autorství textu na základě hodnot jedné funkce, která v ideálním případě zůstává neměnná pro jednoho autora, ale mezi různými autory se liší. Druhou a o něco mladší metodou určování autorství je tzv. vícerozměrná analýza. Ta využívá k určení autorství více různých vlastností textu. Pro texty reprezentované souborem vlastností pak lze využít vhodnou funkci k vypočtení nepodobnosti. Texty tak lze vizualizovat jako body v prostoru, přičemž za pravděpodobného autora anonymního textu je označen autor textu, který je anonymnímu textu nejblíže. Třetí a nejmladší skupina metod k určení autorství využívá strojové učení a neurální sítě. Nástup těchto metod byl významným obratem v historii studia určování autorství. Texty autorů jsou reprezentovány jako číselné vektory reprezentující textové vlastnosti, přičemž použitím vhodné metody strojového učení jsou na jejich základě texty klasifikovány jako náležící k určitému autorovi. (Koppel, Schler, & Argamon, 2009)

2.3.1 Analýza invariant

Analýza invariant se zaměřuje na jednu konkrétní vlastnost textové statistiky, jejíž hodnota zůstává více či méně neměnná pro jednoho konkrétního autora, avšak je proměnlivá mezi více různými autory. Základní myšlenkou analýzy invariant tedy je, že anonymní text bude sdílet podobnou hodnotu zvolené invarianty s jinými texty téhož autora, a tedy porovnáním oné invarianty u anonymního textu a textů potenciálních autorů můžeme determinovat skutečného autora anonymního textu. Pro tento účel byly

navrženu různé statistické vlastnosti textu. Mezi ně se řadily kupříkladu textové vlastnosti jako průměrná délka slov, průměrný počet slov ve větě, poměr typů a tokenů či počet slov se specifickou frekvencí či různá měřítka bohatosti slovníku jako například Yuleho K nebo Sichelovo S. Tyto navrhované invarianty se však neukázaly jako dostatečně spolehlivé samy o sobě, aby se na jejich základě dalo bezpečně autorství určit. (Koppel, Schler, & Argamon, 2009)

2.3.2 Multivariační analýza

Multivariační analýza je přístup, který k určení autorství využívá většího množství vlastností. Těmi mohou být například frekvence určitého výběru slov nebo jiné vyčíslitelné vlastnosti textu. Analyzované texty jsou reprezentovány vektory těchto vlastností. Následně je na tyto vektory aplikována vhodná funkce, pomocí které je dimenzionalita těchto vektorů zredukována tak, aby je bylo možné vizualizovat jako body v dvojrozměrném prostoru. Pokud je mezi známými odlišnými autory dostatečná vzdálenost, pak je za autora anonymního textu označen autor toho textu, který je anonymnímu textu nejblíže. Mezi metody využívající vícerozměrnou analýzu se řadí například Burrowsova Delta, jež s různými variacemi využita na širokou řadu problémů určování autorství. (Koppel, Schler, & Argamon, 2009)

2.3.3 Přístupy se strojovým učením

Metody využívající strojové učení patří k těm nejmladším přístupům k určování autorství textu. Schopnost strojového učení, nalézt hranice mezi texty a klasifikovat je, výrazně překračuje schopnosti funkcí minimalizujících vzdálenost. Jako vstupní data pro strojové učení lze využít různých vyčíslitelných vlastností textu. Schopnost strojového učení spolehlivě určit autorství bude výrazně záviset jak na výběru a kvalitě vstupních dat, tak i na zvolené metodě učení. (Koppel, Schler, & Argamon, 2009)

2.4. Varianty problému určování autorství textu

Při určování autorství textu je možné se setkat s různými variacemi tohoto problému. V ideálních podmínkách máme pro určení autorství k dispozici malou uzavřenou skupinu potenciálních autorů a velice rozsáhlý autorský materiál dostupný k analýze. V praxi se však s takto ideálními podmínkami téměř nikdy nesetkáme. Rozdílné typy problémů určování autorství mají svá specifika a limitace. Hlavními faktory pro

vymezení konkrétního typu problému určení autorství jsou počet potencionálních autorů textu a rozsah textů, které můžeme analyzovat. Skrze tato dvě základní kritéria můžeme vymezit několik základních typů problémů, se kterými se při určování autorství můžeme setkat. (Koppel, Schler, & Argamon, 2009)

Ideální typ problému zahrnuje relativně malou uzavřenou skupinu potenciálních autorů. Naším úkolem zde je vybrat z této skupiny skutečného autora anonymního textu. Docílit toho můžeme pomocí měření míry odlišnosti vlastností ověřených textů autorů od anonymního textu. O míře odlišnosti zde můžeme uvažovat jako o vzdálenosti, přičemž autor, který je anonymnímu textu „nejblíže", je označen za pravděpodobného autora. S tímto typem problému se v praxi nesetkáváme příliš často, jelikož v reálných podmínkách máme uzavřenou malou skupinu kandidátů jen málokdy. (Koppel, Schler, & Argamon, 2009)

Dalším typem je tzv. verifikační problém. Uvažujeme anonymní text a pouze jednoho potencionálního autora. Cílem analýzy tohoto anonymního textu je podat odpověď na otázku, zda onen kandidát je, anebo není autorem anonymního textu. Tento problém lze řešit porovnáním míry rozdílnosti vlastností jiných textů uvažovaného kandidáta a anonymního textu, avšak pouze za předpokladu, že je délka anonymního textu dostatečná. (Koppel, Schler, & Argamon, 2009)

Následující problém nazýváme typem „Jehla v kupce sena". Určujeme autorství anonymního textu pro širokou skupinu potencionálních autorů, přičemž jejich počet může být v řádu tisíců a délky textů, které máme k dispozici pro analýzu, mohou být nedostačující. V takovém případě je velice obtížné určit autora textu spolehlivě. Cílem analýzy je potom spíše snaha o redukci počtu potenciálních kandidátů než o spojení anonymního textu s jedním konkrétním kandidátem. (Koppel, Schler, & Argamon, 2009)

Poslední typ nazýváme „profilující" problém. Představuje situaci, kdy nemáme vůbec žádného potenciálního autora. V takovém případě bude naším cílem pouze profilování autora anonymního textu. Snažíme se získat co možná nejvíce užitečných informací o autorovi textu. Cílem profilování může kupříkladu být získání informace o pohlaví, věku, mateřském jazyku či osobnostních vlastnostech neznámého autora. (Koppel, Schler, & Argamon, 2009)

3. Praktická část

3.1. Úvod praktické části

V rámci praktické části této práce se budeme zabývat disperzí slov v textu ve vztahu k určení autorství textu. V rámci tohoto výzkumu budeme pracovat se souborem textů, jejichž autorství je známé. Nebudeme se tedy pokoušet o určení autorství textu, jehož autorství je sporné, nýbrž budeme posuzovat, jak úspěšně dokáže námi navržená metoda odlišit od sebe navzájem texty jiných autorů a zároveň identifikovat texty stejného autora jako blízké sobě navzájem. Pro odlišení autorství textu se zaměříme na vztah disperze slov, tedy způsobu rozmístění slov v rámci textu, a frekvence výskytu slov. Pokusíme se o rozdělení lexika do několika různých tříd dle povahy jeho disperze a zároveň o posouzení vlastností lexika v jednotlivých třídách pro přenášení informace o distinktivním autorském stylu. Zvláštní pozornost pak budeme věnovat lexiku s nízkou frekvencí v opozici k běžně užívaným metodám určování autorství, které se ve své analýze soustředí zejména na vysoce frekventovaná funkční slova.

3.1.1 Hypotéza

Hypotézu, na níž zakládáme praktickou část této práce, lze shrnout jako následující: Předpokládáme, že rozmístění jednotlivých výskytů daného typu v rámci delšího textu není náhodné, ale obsahuje hodnotnou informaci, která může souviset s autorským stylem, žánrem či tématem textu. Domníváme se, že studium vlastností disperze daného typu uvnitř daného textu nám může poskytnout informace nad rámec poznatků, které bychom získali studiem pouze frekvence výskytů. Zároveň se domníváme, že skrze analýzu disperze u lexika s nižší frekvencí můžeme získat poznatky, které by nebylo možné získat pomocí frekvenční analýzy. Naším předpokladem je, že tato informace má potenciál být užitečnou pro určení autorství textu. Cílem této práce je tedy prozkoumat hypotézu o možnostech využití disperze jako stylometrické metriky. Naše hypotéza vychází z výzkumu Dana Faltýnka, Vladimíra Matlacha a dalších výzkumníků z Univerzity Palackého v Olomouci, kteří se zabývají vlastnostmi nízkofrekventovaného lexika opakovaně se vyskytujícího v rámci textu. Pravidelně opakované nízkofrekventované lexikum nazývají termínem „superhapax“. Tento typ lexika byl shledán jako distinktivní pro specifický autorský styl. Domníváme se proto,

že studiem disperze slov v textech můžeme na tuto práci navázat a rozšířit tak poznatky v této oblasti výzkumu.

3.1.2 Nástroje

K ověření naší hypotézy a ke studiu vlastností disperze u lexika provedeme řadu různých měření na vybraných autorských textech z korpusu. Tyto vybrané texty uložené ve formátu „plain text“ zpracujeme pomocí řady vlastních programovacích scriptů. Pro vytvoření těchto scriptů použijeme programovací jazyky Python a R. Pro zpracování textů budeme používat převážně scripty užívající programovací jazyk Python, který je pro tento typ úlohy vhodnější. Texty zpracované pomocí těchto scriptů budou poté uloženy v podobě datových tabulek. Pro práci s datovými tabulkami použijeme scripty používající programovací jazyk R. Jednou z předností programovacího jazyka R je snadná práce s datovými tabulkami. Díky tomu je programovací jazyk R vhodným nástrojem pro tuto část našeho výzkumu.

3.1.3 Metodika

V rámci našeho výzkumu budeme postupovat následovně. Nejprve se budeme zabývat předzpracováním textu, které budeme používat pro zpracování textů našeho korpusu. Následně již zpracované texty uložíme v podobě datových tabulek. Tyto datové tabulky zpracujeme dále tak, abychom pro každý typ stanovili příslušnost k jedné z definovaných disperzních tříd. Dále se budeme zabývat vlastnostmi a množstvím lexika obsaženého v jednotlivých disperzních třídách tak, abychom získali užitečné poznatky o vlastnostech jednotlivých tříd. V závěrečné fázi získáme pro výběr textů z korpusu výpočet vzájemných nepodobností, na jehož základě vytvoříme nepodobnostní matice. Na vzniklé matice poté použijeme metody vícerozměrného škálování a hierarchického shlukování. Tuto metodu budeme aplikovat na různé výběry lexika takovým způsobem, abychom zjistili, jaké lexikum definované pomocí jeho frekvence výskytu a způsobu disperze je nevhodnější k určování autorství textu.

3.1.4 Korpus

Zdrojem textů pro náš výzkum je referenční korpus pro studium určování autorství, který ve své práci užívá Stefan Evert a jeho kolegové. (Evert, a další, 2017). Texty jsou veřejně dostupné na https://github.com/cophi-wue/refcor. (Schöch, 2017) Korpus

představuje tři sbírky děl určených pro výzkum stylometrického určování autorství. Každá ze sbírek obsahuje 75 děl 25 různých autorů, přičemž každý z nich je reprezentován třemi díly. Sbírky obsahují texty v němčině, angličtině a francouzštině. (Schöch, 2017). Pro potřeby našeho výzkumu se naše analýza omezuje pouze na menší výběr textů z korpusu. Tento výběr omezujeme pouze na texty v anglickém jazyce. V rámci této práce pracujeme pouze s 39 texty pro celkem 13 autorů. Nevyužíváme tedy zdrojový korpus v plné šíři. Datum publikace textů obsažených ve sbírce anglických textů se pohybuje mezi lety 1838 – 1921. Původním zdrojem anglických textů obsažených v korpusu je projekt Gutenberg. Jedná se o online knihovnu poskytující přístup k volně dostupným e-knihám, pro které již vypršela platnost autorských práv. (Project Gutenberg, 2024)

3.1.5 Termíny a pojmy

Zde uvedeme některé termíny a pojmy, které budeme v rámci této práce často užívat. Termín „token“ značí konkrétní realizaci textové jednotky, kterou je v našem případě slovo, v daném textu. Termín „typ“ (anglicky „type“) zde označuje unikátní textovou jednotku, tedy slovo, v rámci textu. Pro jejich vztah tedy platí, že pro každý „typ“ existuje jeden nebo více „tokenů“. Token lze tedy chápat jako konkrétní výskyt slova náležející danému typu. Disperzi slova v textu chápeme jako vztah rozdílů v pořadí jednotlivých tokenů příslušejících stejnému typu v rámci textu o určité délce měřené v počtu tokenů. Disperzi určitého slova tedy chápeme jako vztah rozmístění tokenů náležících nějakému typu, přičemž každý typ má svoji vlastní disperzi. Text chápeme jako seznam tokenů seřazených v určitém pořadí. Každému jednotlivému tokenu přísluší index označující pořadí, ve kterém se v daném textu vyskytuje. Index každého tokenu je unikátní a žádné dva tokeny nemohou sdílet stejný index. Jako vzdálenost dvou tokenů chápeme rozdíl v jejich pořadí. Pro každý typ slova máme množinu jeho tokenů, kdy každému tokenu přísluší index značící jeho pořadí v textu. Vztahy mezi vzdálenostmi tokenů příslušících stejnému typu tak nesou informaci o tom, jakým způsobem jsou tokeny daného typu v textu rozmístěné. Disperzi daného typu v určitém textu tedy můžeme reprezentovat seznamem indexů.

3.2. Analýza disperze slov měřením frekvencí v úsecích

K analýze rozmístění jednotlivých typů uvnitř daného textu využijeme metodu spočívající v rozdělení textu do úseků o shodné délce v počtu tokenů. Pro každý typ získáme jeho frekvenci v každém ze zmíněných úseků. Výsledkem zpracování analyzovaného textu bude datová tabulka obsahující pro všechny typy v daném textu jejich frekvenci, a to v každém z úseků daného textu. Délka těchto úseků a jejich celkový počet jsou parametry, které je nutné určit tak, aby odpovídaly vlastnostem textu a cíli analýzy. Níže je uveden příklad prvních deseti řádků takto vytvořené tabulky.

Tabulka 1: Příklad prvních deseti řádků tabulky

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
forgot	0	0	1	0	0	1	1	0	0	0
proper	1	1	0	4	1	2	1	3	1	1
doom	0	0	0	0	0	1	0	0	0	0
captivated	0	0	0	0	1	0	0	0	0	0
promises	0	1	0	0	0	0	0	0	1	0
lie	1	2	1	0	0	1	1	1	0	0
provided	0	0	0	0	1	0	0	1	0	0
agreement	0	0	0	0	0	1	0	0	0	0
increase	0	0	0	0	0	0	0	0	0	1
parole	0	0	1	0	0	0	0	0	0	0

Získaná tabulka nám umožňuje nahlédnout na způsob, jakým jsou dané typy slov v textu rozprostřeny. Pro hlubší analýzu aplikujeme na tabulku filtry, abychom získali pouze takové typy, jejichž způsob disperze splňuje požadované podmínky.

3.2.1 Předzpracování textu

Prvním krokem je načtení textového řetězce z textového souboru. Následně na tento textový řetězec aplikujeme tzv. tokenizační funkci, pomocí které rozdělíme textový řetězec na jednotlivá slova, tzv. tokeny. Na základě toho, jakou tokenizační funkci zvolíme, můžeme získat mírně odlišný soubor slov. V této práci použijeme pro zpracování textů tokenizační funkci využívající regulárního výrazu k oddělení slov pomocí netisknutelných znaků. Tento regulární výraz má tuto podobu: "\W+". Jedná se o jednu z nejjednodušších metod tokenizace textu. Volba právě této tokenizační funkce

je motivovaná zachováním co největší míry jednoduchosti a transparentnosti. V některých málo častých případech však aplikací této tokenizační funkce získáme nežádoucí tokeny, které nejsou skutečnými nebo úplnými slovy. Tyto chyby v tokenizaci jsou poměrně vzácné, a proto je považujeme za zanedbatelné.

V dalším kroku převedeme všechna velká písmena na písmena malá. Tento zásah do podoby slov zajistí, že dvě realizace stejného slova lišící se pouze v použití kapitálních písmen budou považována za dva tokeny příslušející stejnému typu. Tento krok může mít i nežádoucí dopad. Například obsahuje-li zpracovávaný text vlastní jména, pro která zároveň existuje substantivum nesoucí jiný význam. Při převedení velkých písmen na malá dojde k odstranění rozdílu a potenciální ztrátě informace. Aplikace i vynechání tohoto kroku může mít jak žádoucí tak i nežádoucí dopady na výsledek. Domníváme se, že pro tuto práci je využití tohoto kroku žádoucí, a proto tento krok aplikujeme.

Následujícím krokem je úprava délky všech analyzovaných textů. Všechny texty, které analyzujeme, upravíme na stejnou délku tím, že si pro analýzu ponecháme pouze prvních n tokenů od začátku textu, kdy n je námi zvolená délka textu.

3.2.2 Dělení textu na úseky

Dalším krokem je rozdělení textu na úseky o shodném počtu tokenů. Pro každý typ vypočteme jeho frekvenci v každém z úseků. Délku těchto úseků je potřebné vhodně zvolit na základě délky analyzovaného textu a cílů analýzy. Zvolení příliš krátkých úseků by mělo za následek vytvoření nadměrně dlouhé datové tabulky, jejíž analýza by byla velice náročná a její zpracování by bylo výpočetně náročné. Naopak volba příliš dlouhých úseků by vyústila v datovou tabulku, která by poskytovala minimum relevantní informace o rozmístění slov v analyzovaném textu. Pro naši analýzu jsme jako přiměřenou délku úseků zvolili 3000 – 5000 tokenů v závislosti na délce textu. Volíme texty dlouhé 30000 – 50000 tokenů. Délku úseků vybíráme takovým způsobem, aby všechny úseky byly stejně dlouhé, kupříkladu tedy úseky o délce 5000 tokenů pro texty o délce 50000 tokenů, které rozdělí text na 10 stejně dlouhých úseků.

3.2.3 Filtrování datové tabulky

Vyhledávání slov, jejichž disperze v textu odpovídá určitým vlastnostem, by bez dalších kroků bylo poměrně obtížné. Z toho důvodu doplníme tabulku od další vlastnosti, které budou shrnovat některé vlastnosti daných typů. Tyto další vlastnosti nám usnadní filtrování tabulky. Za užitečné považujeme zejména dvě následující vlastnosti.

1) Celková frekvence typu v textu

2) Počet úseků s nulovou frekvencí

Celkovou frekvenci snadno získáme součtem frekvencí ve všech úsecích pro daný typ. Pomocí celkové frekvence můžeme odfiltrovat slova, jejichž frekvence je příliš nízká, anebo naopak příliš vysoká. U slov s velmi nízkou frekvencí nemá význam disperzi analyzovat. To zahrnuje zejména hapax legomena, dis legomena a tris legomena. Ta zároveň představují signifikantní část všech typů v textu. Slova s výrazně vysokou frekvencí zahrnují převážně funkční slova. V případě, že se analýza na funkční slova nezaměřuje, může být výhodné je také odfiltrovat.

Počet úseků s nulovou frekvencí získáme spočítáme-li v kolika úsecích pro daný typ se frekvence rovná nule. Tato vlastnost nám umožňuje identifikovat slova koncentrující se do shluků. Pokud je počet úseků s nulovou frekvencí blízký celkovému počtu úseků, pak budou všechny výskyty shromážděné jen malém množství úseků. Pokud je tato vlastnost naopak velice nízká či rovná nule, pak je slovo obsaženo ve většině či všech úsecích alespoň jedenkrát.

3.3. Klasifikace lexika do tříd na základě disperze v textu

V této části práce se budeme zabývat klasifikací lexika dle jeho disperze v textech. Na základě vlastností typů se pokusíme roztřídit jednotlivé typy do několika tříd. Vzhledem k tomu, že základní metrikou pro rozdělení lexika do tříd bude způsob jeho disperze v textu, budeme k těmto třídám referovat užitím termínu „disperzní třídy“. Nejprve definujeme třídy takovým způsobem, aby rozdělovaly typy chovající se v textu ideálním způsobem. Později se budeme zabývat problémy, na které v praxi narazíme

při klasifikaci typů, jejichž disperze ideálním způsobem klasifikovat nelze. Nejprve tedy vymezíme disperzní třídy na základě striktních kritérií, načež následně uvedeme obtíže, které při užití striktního vymezení mohou nastat. Posléze představíme alternativní méně striktní způsob klasifikace, jehož volnější kritéria umožní kompenzovat nedostatky striktního způsobu klasifikace.

3.3.1 Disperzní třídy

První disperzní třídu budou tvořit „hapax legomena“, tedy slova, která se v celém textu objevují pouze jednou. Pro každý jednotlivý typ v této třídě najdeme pouze jediný token, a analýza vztahů pozic tokenů tak není z tohoto důvodu vůbec možná. Proto považujeme za vhodné oddělit hapax legomena od zbytku lexika jakožto samostatnou disperzní třídu. Analýza disperze výskytů slov v textu není proveditelná pro slova, která mají pouze jediný výskyt. Domníváme se však, že zahrnutí těchto slov do samostatné třídy je vhodnější postup, než se jejich výskytem v textu nezabývat vůbec. V tabulce níže jsou uvedeny příklady typů náležejících do této třídy se znázorněním rozložení frekvencí výskytu jejich tokenů v měřených úsecích.

Tabulka 2: Příklady třídy hapax legomen

typ	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
tottering	1	0	0	0	0	0	0	0	0	0
vow	0	0	1	0	0	0	0	0	0	0
assurred	0	1	0	0	0	0	0	0	0	0

Následující třídou jsou slova „s extrémně nízkou frekvencí výskytu“. Typy s frekvencí výskytů odpovídající kupříkladu dvěma nebo třem výskytům je vhodné oddělit od zbytku lexika ze stejného důvodu jako výše zmíněnou třídu Hapax legomen. Na rozdíl od Hapax legomen však nemůžeme pro extrémně nízkou frekvenci výskytů přesně stanovit limit počtu výskytů. Ideální hranice bude pravděpodobně záviset na délce měřených úseků, celkové délce textu a zvoleném počtu měřených úseků. Jelikož nemáme k dispozici žádnou metriku, na jejímž základě bychom mohli stanovit minimální počet výskytů nutný k umožnění analýzy disperze daného typu, nezbývá nám jiná možnost než zvolit tuto hranici arbitrárně na základě našeho úsudku. Zde budeme za extrémně nízkou frekvenci považovat frekvenci, která je nižší než čtyři výskyty. V této kategorii tak budou zahrnuta slova s frekvencí nižší než 4 výskyty, která zároveň nejsou hapax legomeny. V tabulce níže jsou uvedeny příklady typů

náležejících do této třídy se znázorněním rozložení frekvencí výskytu jejich tokenů v měřených úsecích.

Tabulka 3: Příklady třídy extrémně nízké frekvence

typ	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
baby	0	2	0	0	1	0	0	0	0	0
slightest	0	0	0	1	1	0	0	0	0	0
places	0	1	1	0	0	1	0	0	0	0

Další disperzní třídu nazveme označením „shluky“. Půjde o takové typy, jejichž tokeny se koncentrují pouze v jednom z měřených úseků a ve zbytku textu se nevyskytují. Zároveň pro ně platí, že jejich celková frekvence je vyšší nebo rovna námi stanovené hranici oddělující třídu extrémně nízké frekvence výskytů. Pro zvolenou hranici minimálně 4 výskytů tedy bude platit, že za patřící do třídy shluků budeme považovat takové typy, které se vyskytují pouze v jednom z měřených úseků a jejichž celková frekvence je vyšší nebo rovna čtyřem. V tabulce níže jsou uvedeny příklady typů náležejících do této třídy se znázorněním rozložení frekvencí výskytu jejich tokenů v měřených úsecích.

Tabulka 4: Příklady třídy shluků

typ	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
beech	0	4	0	0	0	0	0	0	0	0
chopper	0	0	0	8	0	0	0	0	0	0
frida	0	0	6	0	0	0	0	0	0	0

Následující disperzní třídu nazveme označením „stabilní disperze“. Půjde o takové typy, které se vyskytují ve všech měřených úsecích. Bude se tedy jednat o slova, která jsou stabilně rozptýlena po celé délce textu. Označení „stabilní disperze“ volíme tedy právě proto, že se jednotlivé výskyty stabilně objevují po celé délce textu. Minimálním požadavkem pro zahrnutí typu do této třídy tedy je, že se daný typ vyskytuje alespoň jednou v každém z měřených úseků daného textu. Zároveň pro ně platí, že jejich celková frekvence musí být vyšší než námi stanovená hranice extrémně nízké frekvence výskytů. V tabulce níže jsou uvedeny příklady typů náležejících do této třídy se znázorněním rozložení frekvencí výskytu jejich tokenů v měřených úsecích.

Tabulka 5: Příklady třídy stabilní disperze

typ	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
holy	11	11	6	3	5	6	2	3	2	14
she	86	81	82	72	72	73	41	43	30	38
looked	3	8	2	6	2	10	7	6	7	7

Poslední disperzní třídu nazveme označením „nestabilní disperze“. Půjde o typy, které se vyskytují ve větším množstvím úseků, avšak v některých úsecích se nevyskytují vůbec. Představují přechodnou třídu mezi třídou shluků a třídou stabilní disperze. Označení „nestabilní disperze“ poukazuje na skutečnost, že jde o typy opakující se v textu nestabilně. Tyto typy jsou rozprostřené zdánlivě chaotickým těžko předvídatelným způsobem. Zároveň pro ně platí podmínka, že jejich celková frekvence je vyšší než námi stanovená hranice extrémně nízké frekvence výskytů. V tabulce níže jsou uvedeny příklady typů náležících do této třídy se znázorněním rozložení frekvencí výskytu jejich tokenů v měřených úsecích.

Tabulka 6: Příklady třídy nestabilní disperze

typ	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
bed	1	0	5	0	0	1	0	0	1	0
try	0	1	0	0	2	0	1	0	2	0
sin	3	3	3	2	0	0	0	0	0	4

Pro rozdělení typů do výše zmíněných tříd používáme dvě vlastnosti typů. Těmi jsou celková frekvence výskytů a počet úseků s nulovým výskytem. Třídy hapax legomen a extrémně nízké frekvence odlišujeme pouze pomocí celkové frekvence výskytů. Třídu shluků, stabilní disperze a nestabilní disperze odlišujeme pomocí kombinace celkové frekvence výskytů a počtu úseků s nulovým výskytem.

Třída hapax legomen a extrémně nízké frekvence sdílejí stejné kritérium pro jejich odlišení. V obou případech jde o typy, jejichž frekvence je příliš nízká k tomu, abychom mohli analyzovat vztahy mezi jejich umístěním v textu. Z toho důvodu lze uvážit zahrnutí disperzní třídy hapax legomen do disperzní třídy extrémně nízké frekvence. Pokud zachováme striktní klasifikaci, budou disperzní třídy hapax legomen a extrémně nízké frekvence definovány následovně:

- Hapax legomena: celková frekvence výskytů se rovná jedné.
- Extrémně nízká frekvence: Celková frekvence výskytů je nižší než stanovená hranice čtyř výskytů, a zároveň je vyšší než jedna.

Odlišení tříd pomocí počtu úseků s nulovou frekvencí je závislé na počtu úseků, na které je text rozdělen. Pro text dělený na právě 10 úseků jsou hranice pro odlišení tříd pomocí nulového výskytu následující:

- Stabilní disperze: počet prázdných úseků je roven nule.
- Nestabilní disperze: počet prázdný úseků se nachází v intervalu 8 až 1.
- Shluky: počet prázdných úseků je roven devíti.

3.3.2 Problematické aspekty klasifikace disperzních tříd

Klasifikace do tříd pomocí měření počtu prázdných úseků má některé problematické aspekty. Výše uvedený způsob klasifikace je vhodný pouze pro typy, které jsou v textu rozloženy ideálním způsobem. V některých případech se jednotlivé tokeny určitého typu vyskytují velice blízko hranice dvou úseků, což může zapříčinit, že budou klasifikovány jako odlišný typ disperze. Striktní kritéria pro klasifikaci postrádají flexibilitu, která by umožnila správně zařadit typy, u kterých nacházíme výskyty v blízkosti hranic úseků. Nyní uvedeme příklady situací, při nichž dochází při použití striktních kritérií k chybnému zařazení.

Jedním z případů této nespolehlivosti je situace, kdy předpokládaný shluk leží na hranici dvou úseků, a je proto chybně klasifikován jako nestabilní disperze. Přestože jsou výskyty daného typu pravděpodobně koncentrované pouze v jedné části textu, skutečnost, že se nacházejí na hranici dvou měřených úseků, způsobí klasifikační chybu. Striktní kritéria pro třídu shluků stanoví, že se výskyty mohou objevit pouze v jednom měřeném úseku, což je v tomto případě příčinou vzniklé chyby. Níže uvedené příklady budou tímto dotčeny.

Tabulka 7: příklady chyb v klasifikaci

typ	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
dagger	0	0	2	3	0	0	0	0	0	0
eleanor	0	0	10	1	0	0	0	0	0	0
glanced	0	0	0	0	0	4	1	0	0	0

U některých shluků také může nastat situace, kdy se všechny výskyty koncentrují v jednom úseku s výjimkou jediného výskyty, který zapříčiní odlišnou klasifikaci. V tomto případě se nejedná o překročení hranice měřeného úseku jako u předchozích příkladů. Většina výskytů daného typu je však stejně jako u předchozích příkladů koncentrovaná v jedné části textu. Změnu klasifikace na základě jediného výskytu v jiné části textu zde můžeme považovat za nežádoucí. Příklady jsou uvedeny v tabulce níže.

Tabulka 8: Příklady chybné klasifikace

typ	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
daughter	0	0	0	1	0	0	0	7	0	0
farewell	0	0	0	0	9	0	0	0	0	1
gratitude	0	0	1	0	0	0	0	0	0	4

Dalším případem nespolehlivé klasifikace je situace, kdy je převážně stabilně opakující se typ klasifikován do třídy nestabilní frekvence z toho důvodu, že výskyty, které by za ideálních podmínek patřily do jednoho úseku, připadnou do úseků vedlejších, čímž vznikne prázdný úsek. Jde tedy o chybu způsobenou stejnou příčinou jako chyby v prvním případě, tedy výskytem tokenů v blízkosti hranic úseků. Efekt je však opačný. Namísto toho, aby vznikl nenulový úsek navíc jako v prvním případě, je naopak vytvořen nulový úsek, který naruší disperzi typy, jež je ve zbytku úseků stabilní. Tyto typy jsou poté klasifikovány jako nestabilní disperze, přestože jinak vykazují disperzi stabilní. Prozkoumáme-li frekvenci v jednotlivých měřených úsecích u těchto typů, můžeme si povšimnout náhle zvýšené frekvence v úsecích, které sousedí se vzniklým úsekem s nulovou frekvencí. Příklady typů zasažených tímto typem chybné klasifikace jsou uvedeny v tabulce níže.

Tabulka 9: příklady chyb v klasifikaci

typ	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
lifted	4	6	4	4	4	5	0	7	4	1
sent	1	1	2	4	0	4	1	2	2	2
nothing	1	0	3	1	2	2	1	2	2	1

3.3.3 Méně striktní klasifikace disperzních tříd

Abychom se vyvarovali výše zmíněných klasifikačních chyb, využijeme méně striktní metodu klasifikace. Menší striktnost spočívá v zavedení určité míry tolerance v limitech počtu prázdných úseků. Například pro shluky nebudeme vyžadovat koncentraci všech výskytů pouze v jediném úseku, ale budeme tolerovat výskyty i ve dvou úsecích. Pro stabilní disperzi obdobně budeme tolerovat výskyt jednoho prázdného úseku. Tento krok nám pomůže vyhnout se zmíněným klasifikačním chybám, avšak může mít za následek jiné druhy nevhodné klasifikace.

Uvažme příklady v tabulce níže. Tyto typ budou méně striktní klasifikací určeny jako shluky, ale vyvstává u nich otázka, zda by neměly být určeny spíše jako nestabilní disperze.

Tabulka 10: chyby u méně striktní klasifikace

typ	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
beans	3	0	3	0	0	0	0	0	0	0
terror	0	0	2	0	3	0	0	0	0	0
cousin	0	3	0	8	0	0	0	0	0	0

Z uvedených příkladů vyplývá, že striktní i méně striktní klasifikace pomocí počtu prázdných úseků mají každá vlastní nevýhody a typ disperze tak na jejich základě nemůže být určen s absolutní spolehlivostí.

V následující části této práce budeme pracovat s využitím méně striktní klasifikace. Texty budeme dělit na právě 10 úseků. Třídu hapax legomen zahrneme do třídy lexika s extrémně nízkou frekvencí. Kritéria pro určení tříd proto budou následující:

- Extrémně nízká frekvence: celková frekvence je nižší než čtyři výskyty.
- Shluky: počet prázdných úseků je v intervalu 8 – 9 úseků. Celková frekvence je vyšší než čtyři výskyty.
- Stabilní disperze: počet prázdných úseků je v intervalu 0 – 1 úseků. Celková frekvence je vyšší než čtyři výskyty.
- Nestabilní disperze: počet prázdných úseků je v intervalu 2 – 7 úseků. Celková frekvence je vyšší než čtyři výskyty.

3.3.4 Poměr typů a tokenů v jednotlivých třídách

V této sekci se budeme zabývat zastoupením slov v disperzních třídách. Srovnáme zastoupení z pohledu počtů typů obsažených ve třídách i z hlediska celkového počtu tokenů, které do tříd spadají. Poměry zastoupení slov budeme zkoumat na výběru textů z korpusu. Použijeme texty s lišící se celkovou délkou textu i délkou měřených úseků, abychom zjistili, jakou roli hraje délka textu a úseků na počty zastoupených typů a tokenů.

Výběr textů zahrnuje celkem 16 různých textů o stejné celkové délce a shodné délce měřených úseků. Texty pocházejí od čtyř různých autorů, přičemž zahrnují 8 různých děl. Každé z děl je rozděleno na dvě části o požadované délce textu. Na každého autora tak připadají čtyři texty vzniklé ze dvou děl. Toto rozdělení výběru nám umožňuje sledovat rozdíly mezi autory i rozdíly mezi texty pocházejícími ze stejných či odlišných děl téhož autora. To nám může pomoci rozeznat, zda rozdíly poukazují spíše na odlišný styl autorů či odlišné téma textu.

Na tabulkách níže jsou uvedeny počty typů a tokenů pro výběr textů o celkové délce 30000 tokenů, které jsou rozdělené na 10 úseků dlouhých 3000 tokenů.

Tabulka 11: Lexikum s extrémně nízkou frekvencí

Text	E.N.F – typy	E.N.F – tokeny
barclay_ladies_part_0	2902	4143
barclay_ladies_part_1	3087	4499
barclay_rosary_part_0	3198	4503
barclay_rosary_part_1	3066	4414
blackmore_erema_part_0	3120	4418
blackmore_erema_part_1	2968	4076
blackmore_lorna_part_0	3162	4417
blackmore_lorna_part_1	3030	4160
braddon_audley_part_0	3130	4461
braddon_audley_part_1	3340	4767
braddon_fortune_part_0	3866	5448
braddon_fortune_part_1	3405	4802
cbronte_jane_part_0	4064	5648
cbronte_jane_part_1	3912	5447
cbronte_shirley_part_0	4530	6279
cbronte_shirley_part_1	4119	5694

Do třídy lexika s extrémně nízkou frekvencí spadá významná část typů, nicméně jejich celková frekvence v počtu tokenů je relativně nízká. Poměr mezi počtem tokenů a typů se pohybuje okolo hodnoty 1.4 tokenů na 1 typ. Třída lexika s extrémně nízkou frekvencí zastupuje přibližně 78 % všech typů v textu, avšak pouze zhruba 16 % všech tokenů v textu.

Tabulka 12: Lexikum tvořící shluky

text	shluky – typy	shluky – tokeny
barclay_ladies_part_0	82	397
barclay_ladies_part_1	68	333
barclay_rosary_part_0	86	422
barclay_rosary_part_1	69	330
blackmore_erema_part_0	43	205
blackmore_erema_part_1	83	395
blackmore_lorna_part_0	51	277
blackmore_lorna_part_1	68	329
braddon_audley_part_0	86	439
braddon_audley_part_1	94	449
braddon_fortune_part_0	64	307
braddon_fortune_part_1	69	337
cbronte_jane_part_0	54	241
cbronte_jane_part_1	89	465
cbronte_shirley_part_0	77	397
cbronte_shirley_part_1	69	358

Shluky jsou nejméně zastoupenou třídou v textu. Tato třída má nejnižší počet typů i tokenů. Poměr počtu tokenů a typů se pohybuje přibližně v rozmezí 4.7 – 5,3 tokenů na 1 typ. Třída shluků představuje jen velmi malou část textu v celkovém počtu tokenů i v počtu zastoupených typů. Třída shluků zastupuje přibližně 2 % všech typů v textu, a přibližně 1 % všech tokenů v textu.

Tabulka 13: lexikum se stabilní disperzí

text	stabilní – typy	stabilní – tokeny
barclay_ladies_part_0	188	17792
barclay_ladies_part_1	195	18739
barclay_rosary_part_0	189	18489
barclay_rosary_part_1	178	18721
blackmore_erema_part_0	197	19726
blackmore_erema_part_1	192	20278
blackmore_lorna_part_0	187	19510
blackmore_lorna_part_1	191	19750
braddon_audley_part_0	178	18569
braddon_audley_part_1	182	18682
braddon_fortune_part_0	177	17863
braddon_fortune_part_1	175	18982
cbronte_jane_part_0	186	17920
cbronte_jane_part_1	164	17965
cbronte_shirley_part_0	156	16980
cbronte_shirley_part_1	162	17476

Třída lexika se stabilní disperzí zahrnuje poměrně nízký počet typů, avšak v ní obsažená slova mají velmi vysokou frekvenci. Do této třídy spadá většina vysoce frekventovaných funkčních slov. Poměr mezi počtem typů a tokenů dosahuje hodnot přibližně 100 tokenů na 1 typ, a je tak výrazně vyšší než poměry u ostatních tříd. Takto vysoký poměr je zřejmě zapříčiněn výskytem vysoce frekventovaných funkčních slov. Třída stabilní disperze zastupuje přibližně 4 % všech typů v textu, a přibližně 62 % všech tokenů v textu.

Tabulka 14: lexikum s nestabilní disperzí

text	nestabilní – typy	nestabilní – tokeny
barclay_ladies_part_0	803	7668
barclay_ladies_part_1	730	6429
barclay_rosary_part_0	730	6586
barclay_rosary_part_1	744	6535
blackmore_erema_part_0	660	5651
blackmore_erema_part_1	587	5251
blackmore_lorna_part_0	670	5796
blackmore_lorna_part_1	669	5761
braddon_audley_part_0	717	6531
braddon_audley_part_1	716	6102
braddon_fortune_part_0	736	6382
braddon_fortune_part_1	727	5879
cbronte_jane_part_0	738	6191
cbronte_jane_part_1	723	6123
cbronte_shirley_part_0	732	6344
cbronte_shirley_part_1	742	6472

Třída lexika s nestabilní disperzí má v porovnání s lexikem se stabilní disperzí několikanásobně vyšší počty typů a zároveň výrazně nižší počet tokenů. Poměr počtu tokenů a typů dosahuje hodnot přibližně 8 – 9 tokenů na 1 typ. Třída nestabilní disperze zastupuje přibližně 16 % všech typů v textu, a přibližně 21 % všech tokenů v textu.

Poměr tokenů a typů zastoupených v různých disperzních třídách se může lišit s měnící se celkovou délkou textu a délkou úseků. V tabulkách uvedených níže jsou uvedena procentuální zastoupení tokenů a typů v disperzních třídách spolu s použitou délkou textu a úseků. Uvedená procenta jsou průměrnými hodnotami pro 16 textů zmíněných výše.

Tabulka 15: Procentuální zastoupení tokenů ve třídách

délka úseku	délka textu	nízká frekvence	shluky	stabilní	nestabilní
5000	50000	13 %	1 %	68 %	19 %
4500	45000	13 %	1 %	67 %	19 %
4000	40000	14 %	1 %	65 %	19 %
3500	35000	15 %	1 %	64 %	20 %
3000	30000	16 %	1 %	62 %	21 %
2500	25000	18 %	1 %	60 %	21 %
2000	20000	19 %	1 %	56 %	23 %

První tabulka zobrazuje procentuální zastoupení tokenů textu v jednotlivých třídách. V tabulce můžeme vidět, že procentuální zastoupení tokenů třídy nízké frekvence je nejvyšší při nejnižší délce textu a se zvyšující délkou textu se snižuje. Zastoupení tokenů třídy shluků zůstává s měnící se délkou textu neměnné. Zastoupení tokenů třídy stabilní disperze je se zvyšuje s rostoucí délkou textu a úseků. Zastoupení tokenů třídy nestabilních se s rostoucí délkou textu a úseků snižuje.

Tabulka 16: Procentuální zastoupení typů ve třídách

délka úseku	délka textu	nízká frekvence	shluky	stabilní	nestabilní
5000	50000	75 %	2 %	5 %	18 %
4500	45000	76 %	2 %	5 %	18 %
4000	40000	76 %	2 %	5 %	17 %
3500	35000	77 %	2 %	4 %	17 %
3000	30000	78 %	2 %	4 %	16 %
2500	25000	79 %	2 %	4 %	16 %
2000	20000	80 %	2 %	4 %	15 %

Ve druhé tabulce je zobrazeno procentuální zastoupení typů v jednotlivých třídách. Můžeme vidět, že zastoupení v třídě extrémně nízké frekvence se snižuje s rostoucí délkou textu. Zastoupení ve třídě shluků se s měnící se délkou textu výrazně nemění. Zastoupení třídy stabilních se s rostoucí délkou zvyšuje, stejně jako u tokenů této třídy, ale pro typy se jedná pouze o velmi malý nárůst. Zastoupení typů třídy nestabilních se s rostoucí délkou textu zvyšuje. Třída nestabilních tak vykazuje opačný trend pro zastoupení typů a tokenů.

Shrneme-li tyto poznatky, můžeme konstatovat, že pro texty s celkovou délkou 20000 tokenů až 50000 tokenů s délkou měřených úseků 2000 až 5000 tokenů je zastoupení ve třídách následující:

- Extrémně nízká frekvence: 13 % – 19 % tokenů, 75 % – 80 % typů
- Shluky: 1 % tokenů, 2 % typů
- Stabilní disperze: 68 % – 56 % tokenů, 5 % – 4 % typů
- Nestabilní disperze: 19 % – 23 % tokenů, 18 % – 15 % typů

Třída extrémně nízké frekvence se vyznačuje vysokým zastoupením v typech, kdy tvoří více jak tři čtvrtiny všech typů v textu. Zastoupení tokenů je však oproti tomu nižší, kdy tokeny této třídy tvoří méně než pětinu všech tokenů textu. S rostoucí délkou textu se zastoupení snižuje.

Třída shluků zaujímá pouze velmi malou část všech tokenů i typů. Změny v celkové délce textu a délce měřených úseků nemají značný vliv na její zastoupení. Zastoupení typů je vyšší než zastoupení tokenů.

Třída stabilní disperze se vyznačuje vysokým zastoupením v počtu tokenů a nízkým zastoupením v počtu typů. S rostoucí celkovou délkou textu a délkou měřených úseků se její zastoupení v tokenech zvyšuje, zatímco zastoupení v typech se mírně zvyšuje.

Třída nestabilní disperze se vyznačuje relativně mírným rozdílem v zastoupení typů a tokenů. S rostoucí celkovou délkou textu se zastoupení tokenů zvyšuje, zatímco zastoupení typů se snižuje.

3.4. Frekvenční distribuce v jednotlivých třídách

V této části se budeme zabývat rozložením frekvencí výskytů typů v rámci jejich disperzních tříd. V předchozí části práce jsme prozkoumali poměry počtu zastoupených typů a tokenů ve třídách. Na základě poměrů typů a tokenů si však nemůžeme udělat ucelenou představu o lexiku zastoupeném v disperzních třídách. Kupříkladu vysoký počet tokenů na typ, jaký můžeme pozorovat u stabilní disperze, by mohl být způsoben výskytem malého počtu typů s extrémně vysokou frekvencí. Abychom tedy získali konkrétnější představu o distribuci frekvencí typů uvnitř disperzních tříd, prozkoumáme distribuce frekvencí jednotlivě pro každou ze tříd. Frekvence budeme měřit na stejném vzorku 16 textů 4 autorů, který jsme již výše zmínili. Použijeme texty upravené na celkovou délku 50000 tokenů rozdělené na měřené úseky po 5000 tokenech. Pro vizualizaci distribuce frekvencí typů v textech použijeme tabulky, jejichž řádky budou představovat jednotlivé texty, zatímco sloupce budou označovat hladiny frekvence. Hodnoty v tabulkách budou představovat počet typů v dané disperzní třídě, jejichž frekvence výskytů odpovídá určité frekvenční hladině. Vzhledem k tomu, že se rozmezí frekvencí u jednotlivých tříd výrazně liší, zvolíme pro každou disperzní třídu odlišné frekvenční hladiny. Rozsahy frekvenčních hladin nebudou pro všechny hladiny shodné, ale budou rozvrženy takovým způsobem, aby byla kognitivní interpretace tabulek co nejsnazší. Frekvence s velkým počtem typů budou rozděleny do více hladin, zatímco frekvence s velmi nízkým počtem typů sloučíme do jedné hladiny. Rozsah hladin frekvence je uveden v záhlaví tabulek. Na tuto skutečnost je nutné upozornit, aby nedošlo k chybné interpretaci hodnot v tabulkách.

3.4.1 Frekvenční distribuce třídy extrémně nízkofrekventovaného lexika

Frekvenční distribuce třídy extrémně nízkofrekventovaného lexika je výrazně omezena definicí této třídy. Kritéria pro zahrnutí typů do této třídy umožňují pouze jeden, dva až maximálně tři výskyty. Distribuce frekvencí je zde tedy velice přehledná.

Tabulka 17: Frekvenční distribuce třídy extrémně nízkofrekventovaného lexika

text	1 výskyt	2 výskyty	3 výskyty
barclay_ladies_part_0	2471	876	451
barclay_ladies_part_1	2208	821	377
barclay_rosary_part_0	2679	871	460
barclay_rosary_part_1	2511	834	400
blackmore_erema_part_0	2832	817	403
blackmore_erema_part_1	2698	787	364
blackmore_lorna_part_0	2781	829	409
blackmore_lorna_part_1	2680	781	425
braddon_audley_part_0	2655	965	458
braddon_audley_part_1	3019	940	465
braddon_fortune_part_0	3200	1060	548
braddon_fortune_part_1	3294	1040	548
cbronte_jane_part_0	3396	1153	568
cbronte_jane_part_1	3454	1141	527
cbronte_shirley_part_0	3923	1268	600
cbronte_shirley_part_1	3559	1117	553
průměr	2960	956,25	472,25

Nejvíce zastoupené typy jsou hapax legomena, které tvoří zhruba polovinu typů obsažených v této třídě. Čím je frekvence typů nižší, tím více jejich výskytů najdeme.

3.4.2 Frekvenční distribuce třídy shluků

Třída shluků je nejméně zastoupenou disperzní třídou. Slova, která spadají do této třídy, se mohou vyskytovat pouze ve dvou měřených úsecích. Průměrný počet tokenů na typ se pohybuje zhruba okolo pěti tokenů na typ. Z definice třídy také vyplývá, že frekvence nesmí být nižší než čtyři výskyty. Abychom mohli posoudit distribuci

frekvencí v této třídě, zvolíme pro sloupce tabulky nízké hodnoty frekvencí. Vyšší frekvence zahrneme do posledních sloupců tabulky.

Tabulka 18: Frekvenční distribuce třídy shluků

text	4	5	6	7	8	9	10 - 15	>15
barclay_ladies_part_0	100	7	6	3	2	3	6	0
barclay_ladies_part_1	66	11	4	0	0	1	0	0
barclay_rosary_part_0	113	8	5	4	2	4	2	0
barclay_rosary_part_1	79	13	6	2	1	0	0	0
blackmore_erema_part_0	74	7	1	1	1	0	1	3
blackmore_erema_part_1	63	8	6	5	0	0	2	2
blackmore_lorna_part_0	53	10	3	4	3	1	3	7
blackmore_lorna_part_1	80	16	8	4	7	0	1	2
braddon_audley_part_0	78	14	6	3	0	1	2	0
braddon_audley_part_1	100	19	2	4	1	2	5	3
braddon_fortune_part_0	78	12	3	1	3	2	0	2
braddon_fortune_part_1	104	8	4	2	2	3	0	2
cbronte_jane_part_0	87	8	5	4	1	0	2	3
cbronte_jane_part_1	84	8	4	0	1	2	4	2
cbronte_shirley_part_0	80	10	3	3	1	3	5	0
cbronte_shirley_part_1	73	5	3	2	0	2	1	4
průměr	82	10,25	4,3125	2,625	1,5625	1,5	2,125	1,875

V tabulce můžeme vidět, že výrazná část typů ve třídě shluků má frekvenci čtyř výskytů. Typy s frekvencí vyšší než čtyři výskyty tvoří zhruba pětinu typů třídy shluků. Hodnota čtyř výskytů je minimální hodnotou, které v souladu s definicí mohou typy dosáhnout. V případě, že bychom stanovili hranici minimální frekvence vyšší, přišli bychom o významnou část typů v této třídě. Můžeme tedy konstatovat, že mezi shluky běžně nalezneme pouze slova s nízkými frekvencemi výskytu, mezi nimiž jen malá skupina slov bude mít frekvenci vyšší než 10 výskytů. Jelikož jsme pro vytvoření této tabulky použili texty dělené na 10 úseků po 5000 tokenech, můžeme předpokládat, že pro nižší délky úseků pravděpodobně budou počty výskytů ještě nižší.

K lepšímu pochopení distribuce frekvencí v této třídě nám může posloužit náhled na základní statistické vlastnosti. V tabulce níže je uvedeno minimum, první kvartil, medián, průměr, třetí kvartil a maximum pro každý z šestnácti analyzovaných textů.

Tabulka 19: Základní statistiky pro frekvence třídy shluků

text	minimum,	1. kvartil	medián	průměr	3. kvartil	maximum,
barclay_ladies_part_0	4	4	4	5,18	5	15
barclay_ladies_part_1	4	4	4	4,73	5	10
barclay_rosary_part_0	4	4	4	4,90	5	14
barclay_rosary_part_1	4	4	4	4,76	5	9
blackmore_erema_part_0	4	4	4	5,69	5	54
blackmore_erema_part_1	4	4	4	5,40	6	21
blackmore_lorna_part_0	4	4	4	6,54	6,25	34
blackmore_lorna_part_1	4	4	5	5,44	6	19
braddon_audley_part_0	4	4	4	4,98	5,25	15
braddon_audley_part_1	4	4	4	5,38	6	19
braddon_fortune_part_0	4	4	4	5,48	5	45
braddon_fortune_part_1	4	4	4	5,28	5	51
cbronte_jane_part_0	4	4	4	5,35	5	32
cbronte_jane_part_1	4	4	4	5,30	5	28
cbronte_shirley_part_0	4	4	4	5,19	5	15
cbronte_shirley_part_1	4	4	4	5,62	5	26

Hodnota minima, mediánu a prvního kvartilu dosahuje u všech analyzovaných textů čtyř výskytů, což je nejnižší hodnota umožněná definicí této třídy. Průměr se nachází pouze mírně nad mediánem. Hodnota maxima se mezi různými texty výrazně liší. Tyto statistiky potvrzují, že třída shluků je tvořena především velmi nízkými frekvencemi.

3.4.3 Frekvenční distribuce třídy stabilní disperze

Třída stabilní disperze, jak už bylo uvedeno dříve, zahrnuje slova vyskytující se v každém z měřených úseků alespoň jednou, přičemž je tolerován jeden prázdný úsek. Pro tuto třídu je typické velké zastoupení tokenů a malý počet typů. Tato třída zahrnuje vysoce frekventovaná funkční slova i méně frekventovaná slova rozprostřená po celé délce textu. Abychom prozkoumali slova obsažené v této třídě důkladněji, prostudujeme vztah celkové frekvence typů a počtu typů obsažených v této třídě. V tabulce níže jsou zobrazeny počty typů nacházející se v různých hladinách frekvence.

Tabulka 20: Frekvenční distribuce třídy stabilní disperze

text	9-50	51-100	101-150	151-200	201-250	251-1000	>1000
barclay_ladies_part_0	168	64	19	11	9	21	4
barclay_ladies_part_1	169	69	18	9	7	26	4
barclay_rosary_part_0	156	66	19	10	6	25	4
barclay_rosary_part_1	147	73	26	12	5	27	5
blackmore_erema_part_0	131	78	28	12	7	21	6
blackmore_erema_part_1	146	68	29	17	7	23	6
blackmore_lorna_part_0	152	57	25	17	8	21	6
blackmore_lorna_part_1	138	62	32	13	5	24	5
braddon_audley_part_0	156	60	29	9	6	23	5
braddon_audley_part_1	129	63	17	8	7	26	5
braddon_fortune_part_0	138	61	15	16	2	24	5
braddon_fortune_part_1	145	60	16	15	8	23	5
cbronte_jane_part_0	166	49	22	13	5	18	6
cbronte_jane_part_1	135	62	16	19	1	22	6
cbronte_shirley_part_0	141	49	15	13	6	22	5
cbronte_shirley_part_1	157	55	17	13	8	21	5
průměr	148,375	62,25	21,4375	12,9375	6,0625	22,9375	5,125

Z této tabulky je patrné, že třída stabilní disperze zahrnuje větší množství typů s relativně nižší frekvencí a malé množství s velmi vysokou frekvencí. Povšimněme si, že poslední dva sloupce tabulky mají výrazně vyšší rozpětí než zbytek sloupců. Vysoké frekvence jsme zahrnuli do dvou posledních sloupců, a to kvůli usnadnění interpretace tabulky. Poslední řádek tabulky představuje průměrnou hodnotu pro uvedené texty. Typy zahrnuté do třídy stabilní disperze tedy obsahují většinově slova o frekvenci nižší než 100 výskytů a pouze malé množství typů s velmi vysokou frekvencí.

K lepšímu pochopení distribuce frekvencí v této třídě nám může posloužit náhled na základní statistické vlastnosti. V tabulce níže je uvedeno minimum, první kvartil, medián, průměr, třetí kvartil a maximum pro každý z šestnácti analyzovaných textů.

Tabulka 21: Základní statistiky pro frekvence třídy stabilní disperze

text	minimum.	1. kvartil	Medián	Průměr	3. kvartil	maximum.
barclay_ladies_part_0	11	27	42,50	111,17	89,25	3832
barclay_ladies_part_1	10	27	43	113,07	84,50	3637
barclay_rosary_part_0	13	29	47	118,02	90	2485
barclay_rosary_part_1	12	28,50	51	119,66	103	2317
blackmore_erema_part_0	10	30	56	125,89	102	2153
blackmore_erema_part_1	10	27	51	123,81	116,25	1903
blackmore_lorna_part_0	12	28	47	123,41	108	2354
blackmore_lorna_part_1	10	28	52	126,81	107	2191
braddon_audley_part_0	14	29	47	118,48	99,50	2823
braddon_audley_part_1	12	27	49	129,94	98,50	2754
braddon_fortune_part_0	11	27	48	124,32	96	2462
braddon_fortune_part_1	11	27	46,50	123,64	96,50	2402
cbronte_jane_part_0	11	26	40	115,97	85	2319
cbronte_jane_part_1	12	28	48	126,20	98	2150
cbronte_shirley_part_0	13	28,50	45	124,14	97,50	2203
cbronte_shirley_part_1	9	27	43	118,24	98	1967

Hodnota mediánu u stabilní disperze se v analyzovaných textech pohybuje zhruba mezi 40 až 50 výskyty. Hodnota průměru je výrazně vyšší. Pohybuje se zhruba mezi 110 až 130 výskyty. Tento rozdíl je zřejmě způsoben vlivem menšího množství slov s velmi vysokou frekvencí. Hodnota maxima se mezi různými texty může značně lišit. Pohybuje se od zhruba 2000 výskytů až po téměř 4000 výskytů. Nejnižší frekvence umožněná definicí je 9 výskytů, nicméně hodnota minima v analyzovaný textech se pohybuje mezi 9 až 14 výskyty.

3.4.4 Frekvenční distribuce třídy nestabilní disperze

Třída nestabilní disperze zahrnuje slova, která se vyskytují minimálně ve třech úsecích a maximálně v osmi úsecích. Minimální frekvence, kterou mohou typy mít, je rovna 4 výskytům. Výše jsme uvedli, že třída nestabilní disperze představuje 19 % – 23 % tokenů a 18 % – 15 % typů celého textu. Abychom mohli vizualizovat distribuci frekvencí, rozdělíme tabulku na sloupce zvyšující se po deseti výskytech. Do posledního sloupce zahrneme slova, která překračují 50 výskytů.

Tabulka 22: Frekvenční distribuce třídy nestabilní disperze

text	4 – 10	11 – 20	21 – 30	31 – 40	41 – 50	>50
barclay_ladies_part_0	815	250	46	12	0	7
barclay_ladies_part_1	768	261	48	7	3	5
barclay_rosary_part_0	780	222	44	12	5	2
barclay_rosary_part_1	702	214	38	6	1	5
blackmore_erema_part_0	755	180	35	6	1	1
blackmore_erema_part_1	603	171	23	11	6	2
blackmore_lorna_part_0	718	204	35	6	1	3
blackmore_lorna_part_1	696	210	32	5	2	4
braddon_audley_part_0	784	234	30	4	1	3
braddon_audley_part_1	764	242	47	11	4	2
braddon_fortune_part_0	823	214	32	10	3	5
braddon_fortune_part_1	802	194	24	2	6	2
cbronte_jane_part_0	865	196	35	7	2	7
cbronte_jane_part_1	818	203	38	6	0	4
cbronte_shirley_part_0	875	203	46	10	4	5
cbronte_shirley_part_1	860	220	30	3	2	3
průměr	776,75	213,625	36,4375	7,375	2,5625	3,75

V tabulce můžeme vidět, že většina slov ve třídě nestabilní disperze má frekvenci mezi čtyřmi a dvaceti výskyty. Podobně jako u ostatních tříd počet zastoupených typů se zvyšující se frekvencí výskytů prudce klesá. Slova s frekvencí výskytu vyšší než 50 výskytů budou v této třídě spíše výjimkou.

K lepšímu pochopení distribuce frekvencí v této třídě nám může posloužit náhled na základní statistické vlastnosti. Na tabulce níže je uvedeno minimum, první kvartil, medián, průměr, třetí kvartil a maximum pro každý z šestnácti analyzovaných textů.

Tabulka 23: Základní statistiky pro frekvence třídy nestabilní disperze

text	minimum	1. kvartil	medián	průměr	3. kvartil	maximum
barclay_ladies_part_0	4	5	7	9,61	11	215
barclay_ladies_part_1	4	5	7	9,60	12	167
barclay_rosary_part_0	4	5	7	9,17	11	90
barclay_rosary_part_1	4	5	7	9,15	11	135
blackmore_erema_part_0	4	5	6	8,38	10	66
blackmore_erema_part_1	4	5	7	9,22	11	81
blackmore_lorna_part_0	4	5	7	8,78	11	65
blackmore_lorna_part_1	4	5	7	8,92	11	90
braddon_audley_part_0	4	5	7	8,90	11	253
braddon_audley_part_1	4	5	7	9,20	11	72
braddon_fortune_part_0	4	5	7	9,23	10	239
braddon_fortune_part_1	4	5	7	8,44	10	103
cbronte_jane_part_0	4	5	7	8,68	10	104
cbronte_jane_part_1	4	5	6	8,60	10	77
cbronte_shirley_part_0	4	5	6	8,78	10	101
cbronte_shirley_part_1	4	4	6	8,42	10	147

U třídy nestabilní disperze je hodnota průměru a mediánu nepříliš vzdálená od minima. Medián se pohybuje v hodnotách mezi 6 a 7 výskyty, zatímco průměr je mírně vyšší s hodnotou zhruba okolo 9 výskytů. Hodnota maxima je výrazně vzdálená od frekvencí mezi prvním a třetím kvartilem. Maximum se mezi analyzovanými texty značně liší. Dosahuje hodnot zhruba mezi 70 výskyty a 250 výskyty.

3.4.5 Zhodnocení frekvenční distribuce tříd

Shrneme-li poznatky, které jsme získali vhledem do frekvenční distribuce jednotlivých tříd, můžeme konstatovat následující. Třída extrémně nízké frekvence se od ostatních nutně odlišuje na základě kritérií definice, a tak nemá význam ji s ostatními třídami srovnávat. Třída stabilní disperze dosahuje nejvyšších frekvencí, přičemž některé typy dosahují až stovek výskytů v textech, které jsme analyzovali. Většinová část typů stabilní disperze ale dosahuje frekvencí nižších než 100 výskytů. Třída nestabilní disperze oproti tomu málokdy přesáhne frekvenci 50 výskytů. Nejnižší rozdíly frekvencí najdeme u třídy shluků, kde frekvence výskytů jen málokdy přesáhne frekvenci 15 výskytů. Třídy se tedy navzájem výrazně liší v rozsahu frekvencí, které se v nich nacházejí.

3.5. Vztah disperze a sémantiky lexika

V této části prozkoumáme vztah mezi sémantikou slov a jejich disperzní třídou. Předpokládáme, že frekvence a způsob rozmístění slova v textu mají vztah k sémantice lexika a jeho roli v textu. Kupříkladu u funkčních slov bychom očekávali stabilní rozmístění v textu a vysokou frekvenci, naopak u podstatných jmen označujících fyzické objekty bychom mohli očekávat, že se budou vyskytovat pouze v krátkých úsecích textu, pokud však nemají důležitý vztah k centrálnímu tématu daného textu. Pro tuto analýzu budeme postupovat tak, že vybereme čtyři texty různých autorů, pro které uvedeme příklady slov náležících do některé z disperzních tříd s určitou frekvencí výskytů. Porovnáme například, jaká slova dané třídy se objevují s vysokou frekvencí, a jaká v nízkých frekvencích. Tuto analýzu provedeme pro třídy stabilní disperze, nestabilní disperze a třídy shluků. Třídu extrémně nízké frekvence z této analýzy vyřadíme.

Pro výběr příkladů použijeme čtyři texty, které jsme již studovali v rámci výběru šestnácti textů. Každý ze čtyř autorů zastoupených dříve zmíněnými šestnácti texty bude reprezentován pouze jedním svým textem. Použijeme texty o délce 50000 tokenů dělené po úsecích dlouhých 5000 slov. V tomto menším výběru budou zahrnuty tyto texty:

- barclay_ladies_part_0
- blackmore_erema_part_0
- braddon_audley_part_0
- cbronte_jane_part_0

3.5.1 Příklady stabilní disperze

Nejprve nahlédneme na třídu stabilní disperze. Uvedeme příklady slov s nejvyššími frekvencemi, poté výběr slov s frekvencí nižší než 50 výskytů a příklady slov s nejnižší možnou disperzí pro stabilní disperzi, kterou je frekvence devíti výskytů.

V tabulkách níže jsou uvedeny příklady pěti typů s nejvyšší frekvencí ve třídě stabilní disperze pro výběr čtyř textů.

Tabulka 24: Příklady stabilní disperze s vysokou frekvencí výskytů

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
the	416	392	420	316	361	435	397	365	400	330
of	169	150	158	131	128	160	169	142	132	137
and	142	116	156	141	161	120	147	128	155	149
to	116	139	139	156	142	122	130	165	137	153
her	114	106	114	93	101	104	48	77	45	65

Tabulka 25: Příklady stabilní disperze s vysokou frekvencí výskytů

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
the	211	227	202	279	257	220	218	131	208	200
and	196	222	190	210	211	209	198	190	197	213
of	155	178	172	158	152	148	150	157	141	133
to	135	158	140	134	140	168	161	165	150	171
i	110	130	129	121	144	120	71	195	95	116

Tabulka 26: Příklady stabilní disperze s vysokou frekvencí výskytů

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
the	280	265	288	279	275	329	294	304	277	232
and	189	189	156	150	143	143	126	155	150	148
of	132	136	126	142	123	124	140	146	148	107
to	105	132	118	147	147	117	123	128	113	126
a	156	124	152	104	127	114	102	94	125	125

Tabulka 27: Příklady stabilní disperze s vysokou frekvencí výskytů

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
the	243	198	206	301	252	202	239	215	214	249
and	175	176	197	202	153	184	194	202	195	194
i	139	180	210	136	169	132	199	222	155	174
to	152	129	149	122	182	150	147	144	146	132
of	144	108	115	150	143	154	131	103	131	97

Na příkladech můžeme vidět, že na nejvyšších frekvencích stabilní disperze se vyskytují většinou stejná slova. Jedná se o nejfrekventovanější funkční slova daného jazyka. Lze předpokládat, že tato slova budou vždy přítomna v třídě stabilní disperze.

Dále uvedeme příklady pěti typů s nejvyšší frekvencí nižší než 50 výskytů ve třídě stabilní disperze pro každý ze čtyř zvolených textů. Frekvence 50 výskytů odpovídá zhruba hodnotě mediánu frekvencí výskytů u třídy stabilní disperze.

Tabulka 28: Příklady stabilní disperze s frekvencí výskytů v hodnotě mediánu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
first	4	10	7	6	2	2	0	6	8	4
head	4	4	4	3	8	8	5	6	2	5
still	3	3	11	7	9	2	1	3	5	5
left	6	4	9	8	7	6	0	3	4	2
time	7	4	3	4	6	3	2	2	10	8

Tabulka 29: Příklady stabilní disperze s frekvencí výskytů v hodnotě mediánu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
let	8	3	6	3	4	6	3	5	4	7
miss	0	7	3	3	3	2	5	11	6	8
again	5	3	3	6	7	7	7	3	6	1
went	0	1	5	10	5	9	4	3	6	4
water	10	5	1	9	10	3	4	1	1	2

Tabulka 30: Příklady stabilní disperze s frekvencí výskytů v hodnotě mediánu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
lucy	20	0	1	4	3	4	1	9	1	6
here	8	3	7	5	4	7	8	0	2	5
night	1	4	6	3	7	10	7	3	1	7
life	10	7	1	9	5	3	5	3	4	2
much	10	5	6	6	5	3	6	1	2	4

Tabulka 31: Příklady stabilní disperze s frekvencí výskytů v hodnotě mediánu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
made	2	3	6	6	3	4	4	5	7	9
after	2	8	4	5	4	5	9	5	6	1
has	5	2	2	4	5	10	2	3	9	7
half	5	4	3	4	6	4	5	10	2	6
think	5	8	5	1	3	7	1	6	4	7

V příkladech stabilní frekvence s frekvencí nižší než 50 výskytů převažují významová slova, namísto slov funkčních. Také si můžeme povšimnout rozdílů ve frekvencích v jednotlivých úsecích. Přestože se tato slova vyskytují ve všech úsecích nebo naprosté většině úseků, v některých úsecích je frekvence vyšší, než ve zbytku úseků. Slova tedy nejsou rozmístěná rovnoměrně. Mezi příklady se nacházejí různé slovní druhy, aniž by jeden konkrétní slovní druh převažoval.

Následně uvedeme příklady pěti typů s nejnižší frekvencí v třídě stabilní disperze. Minimální frekvence umožněná použitou definicí stabilní disperze je devět výskytů.

Tabulka 32: Příklady stabilní disperze s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
grant	1	1	2	1	0	1	2	1	1	1
kindly	1	1	2	0	1	1	2	1	1	2
every	0	1	2	1	2	2	1	2	1	1
safely	2	0	1	1	1	3	1	1	2	1
whole	1	0	1	2	1	1	2	1	1	3

Tabulka 33: Příklady stabilní disperze s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
age	0	1	1	1	1	1	1	1	2	1
harm	1	1	1	1	1	1	2	2	0	2
bring	1	1	2	0	1	1	1	2	2	1
small	2	1	1	1	1	0	2	1	1	2
between	1	1	2	1	1	0	1	3	1	1

Tabulka 34: Příklady stabilní disperze s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
taking	0	2	2	1	1	2	2	1	1	2
voice	4	1	0	1	1	1	2	1	1	2
ground	1	1	2	1	2	1	1	1	1	3
though	1	1	0	2	1	1	1	5	2	1
around	1	1	1	1	1	2	4	1	3	0

Tabulka 35: Příklady stabilní disperze s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
past	1	1	1	1	2	1	1	1	1	1
question	2	2	1	1	0	1	2	1	3	1
days	1	2	1	0	1	1	4	1	1	2
fell	2	2	1	1	1	2	2	2	1	1
chapter	2	2	1	1	2	2	1	2	1	1

Na nejnižších frekvencích už funkční slova nenacházíme. Namísto toho najdeme směs významových slov zahrnujících substantiva, adjektiva, slovesa i adverbia. Většina těchto slov je v textu rozprostřena rovnoměrně, ačkoliv můžeme najít i takové typy, u kterých se frekvence v konkrétním úseku zvýší. Pro takto nízké frekvence však platí, že pokud by výskyty nebyly rozložené rovnoměrně, nutně by musely vzniknout prázdné úseky, které by znemožnily dané typy klasifikovat jako stabilní disperzi.

Můžeme konstatovat, že u slov zahrnutých do třídy stabilní disperze se projevuje značná variabilita slovních druhů v závislosti na frekvenci výskytů daných slov. Ve vysokých frekvencích budou dominovat funkční slova, zatímco v nižším frekvencích najdeme slova významová.

3.5.2 Příklady nestabilní disperze

Dále se podíváme na příklady typů ze třídy nestabilní disperze. Opět vybereme nejprve příklady typů s nejvyšších frekvencí. Poté zvolíme příklady s frekvencí blízkou úrovni mediánu pro nestabilní disperze, jež se pohybuje okolo hodnoty 7 výskytů. Následně na to vybereme příklady z nejnižších frekvencí, přičemž minimální frekvence umožněná definicí jsou 4 výskyty.

Níže uvedeme 5 příkladů s nejvyšší frekvencí pro třídu nestabilní disperze.

Tabulka 36: Příklady nestabilní disperze s nejvyšší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
bishop	2	0	0	6	1	12	58	52	41	43
antony	36	22	45	1	10	46	0	0	5	0
hugh	0	0	0	10	14	8	2	9	10	35
thy	15	10	19	24	5	4	0	0	2	3
cell	20	9	15	3	12	12	5	2	0	0

Tabulka 37: Příklady nestabilní disperze s nejvyšší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
sawyer	0	5	7	16	5	13	12	3	5	0
river	0	11	2	4	6	5	2	4	0	11
gold	1	2	4	0	8	7	15	0	0	1
large	8	5	3	4	0	2	0	5	3	6
oh	8	0	0	4	3	2	2	6	3	7

Tabulka 38: Příklady nestabilní disperze s nejvyšší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
robert	0	0	26	30	26	42	47	12	38	32
phoebe	0	5	12	1	12	3	0	19	2	18
friend	0	0	14	10	12	12	10	3	3	4
cousin	0	1	8	5	11	0	0	5	16	2
luke	0	5	16	0	0	0	0	7	1	11

Tabulka 39: Příklady nestabilní disperze s nejvyšší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
bessie	14	41	25	6	0	0	0	16	0	2
reed	23	23	28	2	3	4	1	8	0	0
temple	0	0	2	7	21	21	11	0	0	1
helen	0	0	0	0	11	29	22	0	0	0
jane	6	10	13	3	0	9	7	6	1	0

Většina slov s nejvyššími frekvencemi ve třídě nestabilní disperze se skládá ze substantiv, která jsou zároveň vlastními jmény. Vzhledem ke skutečnosti, že analyzované texty jsou žánrem beletrie, domníváme se, že se v tomto případě může jednat o vlastní jména postav hrajících významnější roli v rámci daných děl. Potvrzení této domněnky by však vyžadovalo znalost obsahu děl.

Dále uvedeme příklady slov s frekvencí nižší než 8 výskytů. Hodnota 7 výskytů zhruba odpovídá hodnotě mediánu pro třídu nestabilní disperze v analyzovaných textech.

Tabulka 40: Příklady nestabilní disperze s frekvencí výskytu v hodnotě mediánu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
hell	2	0	0	2	0	0	0	2	1	0
queen	0	0	0	5	0	0	0	1	0	1
command	0	1	2	3	0	0	0	0	1	0
withdrew	1	1	1	0	0	0	2	1	0	1
gaze	1	0	0	1	0	1	0	0	4	0

Tabulka 41: Příklady nestabilní disperze s frekvencí výskytu v hodnotě mediánu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
lie	1	0	1	0	0	2	1	1	1	0
places	1	1	1	1	1	0	0	0	1	1
chief	0	2	0	0	1	0	1	0	2	1
free	2	0	0	1	1	1	0	2	0	0
sent	1	0	1	2	0	1	1	0	0	1

Tabulka 42: Příklady nestabilní disperze s frekvencí výskytu v hodnotě mediánu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
cut	0	0	1	3	3	0	0	0	0	0
easily	0	0	1	1	1	0	0	1	2	1
receive	0	0	0	0	1	0	1	1	0	4
become	2	1	0	2	0	1	0	1	0	0
everything	0	1	0	0	1	1	2	0	2	0

Tabulka 43: Příklady nestabilní disperze s frekvencí výskytu v hodnotě mediánu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
become	2	0	1	0	1	0	2	0	0	1
everything	0	0	0	2	0	0	0	2	3	0
lit	0	0	1	2	0	1	0	1	0	2
tale	1	1	1	0	1	1	0	0	2	0
power	2	0	0	0	0	1	2	1	1	0

U příkladů frekvencí nižších než 8 výskytů nacházíme pouze významová slova. To je však očekávatelné, vzhledem k jejich nízké frekvenci. Najdeme zde převážně substantiva a verba.

Následně prozkoumáme příklady nejnižších frekvencí pro třídu nestabilní disperze.

Tabulka 44: Příklady nestabilní disperze s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
shaking	0	0	2	0	1	1	0	0	0	0
wanted	1	1	1	1	0	0	0	0	0	0
infant	1	0	1	0	0	2	0	0	0	0
mean	0	0	1	0	0	1	1	0	1	0
grass	0	2	0	0	1	0	0	0	0	1

Tabulka 45: Příklady nestabilní disperze s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
proper	0	0	1	0	0	0	2	1	0	0
gazed	1	0	0	1	1	0	1	0	0	0
heads	1	0	0	0	0	1	1	0	0	1
trifle	0	0	0	1	0	0	0	0	2	1
order	0	1	1	0	1	0	0	0	0	1

Tabulka 46: Příklady nestabilní disperze s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
memory	0	1	0	1	0	1	1	0	0	0
lover	1	1	1	0	0	0	0	1	0	0
eating	0	0	0	1	1	0	1	1	0	0
glistening	1	0	0	0	0	0	0	1	1	1
violence	0	0	0	1	0	1	0	0	2	0

Tabulka 47: Příklady nestabilní disperze s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
seek	0	0	1	0	0	0	1	0	1	1
swept	0	1	1	0	0	1	1	0	0	0
hastened	0	0	0	1	0	0	0	1	0	2
sweetest	0	0	1	0	0	1	2	0	0	0
remainder	0	1	0	0	2	1	0	0	0	0

U nejnižších frekvencí třídy nestabilní disperze nacházíme substantiva, verba i adjektiva. Převážně v příkladech vidíme substantiva a verba, podobně jako u příkladů z úrovně mediánu. Jedná se o významová slova. V takto nízkých frekvencích funkční slova nenacházíme.

3.5.3 Příklady shluků

Jako poslední uvedeme příklady třídy shluků. Obdobně jako u předchozích tříd nejprve uvedeme příklady slov s nejvyššími frekvencemi pro tuto třídu. Poté uvedeme příklady nejnižších frekvencí. Frekvence výskytů blízké hodnotám mediánu uvádět nebudeme, jelikož v případě třídy shluků se hodnota mediánu shoduje s hodnotou minima, tedy 4 výskytů, což je zároveň nejnižší hodnota umožněná definicí této třídy.

Níže jsou uvedeny příklady slov s nejvyššími frekvencemi pro třídu shluků.

Tabulka 48: Příklady shluků s nejvyšší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
weeds	0	11	4	0	0	0	0	0	0	0
icon	0	0	0	0	0	0	0	0	13	1
humphry	0	0	0	11	1	0	0	0	0	0
chaplain	0	0	0	0	0	0	8	0	3	0
eleanor	0	0	0	10	1	0	0	0	0	0

Tabulka 49: Příklady shluků s nejvyšší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
major	0	0	0	0	0	0	0	0	29	25
hockin	0	0	0	0	0	0	0	0	18	24
mrs	0	0	0	0	0	0	0	0	4	12
bruntsea	0	0	0	0	0	0	0	0	0	12
swept	0	0	0	6	3	0	0	0	0	0

Tabulka 50: Příklady shluků s nejvyšší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
harry	0	0	0	0	0	0	0	0	12	3
towers	0	0	0	0	0	0	0	0	11	2
morley	0	10	0	0	0	0	0	0	0	0
canaries	0	0	0	2	0	0	6	0	0	0
cabin	3	5	0	0	0	0	0	0	0	0

Tabulka 51: Příklady shluků s nejvyšší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
rochester	0	0	0	0	0	0	0	0	16	16
adele	0	0	0	0	0	0	0	0	11	15
abbot	9	15	0	0	0	0	0	0	0	0
lloyd	0	9	0	0	0	3	0	0	0	0
hay	0	0	0	0	0	0	0	0	2	10

Mezi uvedenými příklady převažují substantiva, z nichž velká část jsou vlastní jména. Jelikož analyzovaná díla jsou žánrem beletrie, lze se domnívat, že se může jednat o jména vystupujících postav. Pro většinu z těchto slov platí, že se vyskytují v rozmezí dvou úseků.

Následně nahlédneme na příklady s nejnižšími frekvencemi výskytu, které jsou zároveň minimem i mediánem distribuce frekvencí v této třídě. Jde tedy o frekvenci výskytu většiny slov v této disperzní třídě.

Tabulka 52: Příklady shluků s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
sounded	0	0	3	0	0	0	0	1	0	0
pillar	0	0	0	1	0	3	0	0	0	0
chuckled	2	0	0	0	0	2	0	0	0	0
hedge	0	4	0	0	0	0	0	0	0	0
oaken	2	2	0	0	0	0	0	0	0	0

Tabulka 53: Příklady shluků s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
toad	0	0	0	0	0	0	4	0	0	0
wonderful	0	0	0	0	1	0	0	3	0	0
penny	0	0	0	0	0	0	0	1	0	3
ranch	0	1	3	0	0	0	0	0	0	0
speech	0	3	0	0	0	1	0	0	0	0

Tabulka 54: Příklady shluků s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
fit	0	3	0	0	0	0	0	0	1	0
introduction	0	0	0	2	2	0	0	0	0	0
land	0	3	0	1	0	0	0	0	0	0
moat	3	1	0	0	0	0	0	0	0	0
knife	0	2	0	0	0	0	0	2	0	0

Tabulka 55: Příklady shluků s nejnižší frekvencí výskytu

	úsek 1	úsek 2	úsek 3	úsek 4	úsek 5	úsek 6	úsek 7	úsek 8	úsek 9	úsek 10
eyebrows	0	0	1	0	0	0	0	0	0	3
millions	0	0	0	0	0	2	0	0	2	0
committee	0	0	0	0	0	0	1	3	0	0
land	0	0	0	0	0	1	0	0	3	0
vous	0	0	0	0	0	0	0	0	2	2

Mezi příklady nejnižších frekvencí třídy shluků nalezneme pouze významová slova. V uvedených příkladech převažují substantiva, méně často verba či jiné slovní druhy. Podíváme-li se na rozložení výskytů v úsecích, můžeme vidět, že pro některé z uvedených příkladů platí, že jejich výskyty neleží v sousedících úsecích. Pro tato

slova se nabízí otázka, zda je vhodné je klasifikovat jako shluky, a to vzhledem k tomu, že jejich výskytů je velmi málo a leží značně vzdálené od sebe navzájem.

3.6. Lexikum sdílené mezi autory

V této části se budeme zabývat vztahy mezi autory a lexikem, které využívají, z pohledu jeho disperze v textu. Dosud jsme vymezili disperzní třídy, do kterých lze lexikum rozřadit, a popsali jsme jeho distribuci uvnitř tříd. Zkoumali jsme počty výskytů v různých disperzních třídách mezi různými autory. Nyní se však zaměříme konkrétně na rozdíly mezi autory, na to jaké typy v různých disperzních třídách využívají.

Nyní použijeme výběr 16 textů čtyř autorů s délkou úseku 5000 tokenů a celkovou délkou 50000 tokenů, který jsme využili už dříve. Pro každou dvojici textů z tohoto výběru změříme počet typů sdílených oběma texty. Seznamy typů, kterými jsou texty reprezentovány, vždy omezíme na typy příslušející k určité disperzní třídě. Tento postup zopakujeme pro různé disperzní třídy. Výsledkem tak bude matice počtů sdílených typů. Cílem této analýzy bude odhalit, zda je lexikum se specifickou disperzí specifické pro autora, či zda je na autorství textu nezávislé. Očekáváme, že texty napsané jedním autorem budou mít mezi sebou navzájem větší množství sdílených typů než spolu s texty jiných autorů.

Jako první disperzní třídu prozkoumáme třídu stabilní disperze. Můžeme očekávat, že vysoce frekventovaná funkční slova budou obsažena ve většině z analyzovaných textů. Významná část této třídy je však tvořena i nižšími frekvencemi, u kterých by se typy specifické pro daného autora či téma textu mohly objevit.

Níže je zobrazena matice počtů sdílených typů ve třídě stabilní disperze pro jednoho autora. V rámci matice jsou porovnány texty i samy se sebou. V takovém případě je mezi nimi úplná shoda.

Tabulka 56: Matice sdílených typů stabilní disperze u textů jednoho autora

	barclay_ladies_part_0	barclay_ladies_part_1	barclay_rosary_part_0	barclay_rosary_part_1
barclay_ladies_part_0	296			
barclay_ladies_part_1	228	302		
barclay_rosary_part_0	199	200	286	
barclay_rosary_part_1	196	200	221	295

Na matici můžeme vidět, že texty stejného autora spolu navzájem sdílí z původních zhruba 300 typů pouze okolo 200 typů. Zároveň můžeme vidět vyšší počet sdílených typů u textů, které pocházejí z různých částí téhož díla. Zde se pravděpodobně projevuje spíše vliv tématu textu.

V matici níže jsou uvedeny počty sdílených typů stabilní disperze textů prvního autora spolu s texty zbylých tří autorů.

Tabulka 57: Matice sdílených typů stabilní disperze s texty jiných autorů

	barclay_ladies_part_0	barclay_ladies_part_1	barclay_rosary_part_0	barclay_rosary_part_1
blackmore_erema_part_0	195	188	202	202
blackmore_erema_part_1	196	189	203	207
blackmore_lorna_part_0	195	189	198	193
blackmore_lorna_part_1	198	191	202	198
braddon_audley_part_0	188	194	207	210
braddon_audley_part_1	172	169	188	194
braddon_fortune_part_0	175	174	198	197
braddon_fortune_part_1	187	187	203	206
cbronte_jane_part_0	195	193	210	212
cbronte_jane_part_1	188	187	197	203
cbronte_shirley_part_0	181	181	193	196
cbronte_shirley_part_1	192	181	199	207

Pozorujeme zde mírně nižší počet sdílených typů než mezi texty stejného autora. V některých případech ale počet sdílených typů dosahuje téměř stejných hodnot, jako počet sdílených typů u textů jednoho autora pocházejících ze různých děl tohoto autora. Na této matici tedy vidíme drobné rozdíly mezi autory, avšak jde o rozdíly poměrně malé. V některých případech sdílejí různí autoři obdobné množství typů se stabilní disperzí, jako sdílejí i mezi jejich vlastními texty.

Dále se zaměříme na lexikum klasifikované do třídy nestabilní disperze. Tato třída obsahuje oproti třídě stabilních větší počty typů, nicméně jejich průměrná frekvence v textu je nižší.

Níže se nachází matice počtů sdílených typů pro typy nestabilní disperze u jednoho autora.

Tabulka 58: Matice sdílených typů nestabilní disperze u textů jednoho autora

	barclay_ladies_part_0	barclay_ladies_part_1	barclay_rosary_part_0	barclay_rosary_part_1
barclay_ladies_part_0	1130			
barclay_ladies_part_1	657	1092		
barclay_rosary_part_0	446	455	1065	
barclay_rosary_part_1	443	433	553	966

Ve třídě nestabilní disperze sdílejí texty jednoho autora přibližně mezi 450 – 500 typy. V celkovém počtu sdílejí více typů, nicméně v poměru k celkovému počtu typů je podobnost mezi texty nižší. Pro texty pocházející ze stejného díla opět vidíme větší počet sdílených vlivů. Opět se zde nejspíše projevuje vliv tématu děl.

Dále následuje matice sdílených typů nestabilní disperze s ostatními autory.

Tabulka 59: Matice sdílených typů nestabilní disperze s texty jiných autorů

	barclay_ladies_part_0	barclay_ladies_part_1	barclay_rosary_part_0	barclay_rosary_part_1
blackmore_erema_part_0	401	363	409	395
blackmore_erema_part_1	326	308	361	346
blackmore_lorna_part_0	373	335	364	330
blackmore_lorna_part_1	433	380	408	387
braddon_audley_part_0	394	387	435	429
braddon_audley_part_1	388	385	419	398
braddon_fortune_part_0	409	395	458	418
braddon_fortune_part_1	412	387	450	416
cbronte_jane_part_0	416	396	466	443
cbronte_jane_part_1	425	423	470	450
cbronte_shirley_part_0	414	402	456	438
cbronte_shirley_part_1	440	397	469	447

Porovnáme-li prvního autora s autory ostatními, zjistíme, že v některých případech je podobnost nižší než pro autora srovnávaného s vlastním textem, avšak u některých autorů je počet sdílených typů obdobný, jako u počtu typů se společným autorem textu. Nacházíme přibližně 300 – 500 sdílených typů mezi každou dvojicí textů. Výsledky jsou tedy podobné těm, které jsme viděli u třídy stabilní disperze.

Následně prozkoumáme třídu shluků. Tato třída se vyznačuje velmi nízkým počtem typů s nízkou frekvencí. Očekáváme, že se budou silně vázat k tématu textu.

Níže je zobrazena matice počtů sdílených typů ve třídě shluků pro jednoho autora.

Tabulka 60: Matice sdílených typů shluků u textů jednoho autora

	barclay_ladies_part_0	barclay_ladies_part_1	barclay_rosary_part_0	barclay_rosary_part_1
barclay_ladies_part_0	127			
barclay_ladies_part_1	5	82		
barclay_rosary_part_0	4	0	138	
barclay_rosary_part_1	2	0	3	101

V matici můžeme vidět, že počet sdílených typů pro třídu shluků pro texty jednoho autora je extrémně nízký, někdy až nulový.

Dále uvedeme matici sdílených typů třídy shluků u jednoho autora s ostatními autory.

Tabulka 61: Matice sdílených typů shluků s texty jiných autorů

	barclay_ladies_part_0	barclay_ladies_part_1	barclay_rosary_part_0	barclay_rosary_part_1
blackmore_erema_part_0	3	2	1	1
blackmore_erema_part_1	2	1	3	3
blackmore_lorna_part_0	2	0	0	2
blackmore_lorna_part_1	2	2	2	1
braddon_audley_part_0	1	2	4	2
braddon_audley_part_1	3	1	5	3
braddon_fortune_part_0	2	2	4	2
braddon_fortune_part_1	0	4	2	2
cbronte_jane_part_0	4	3	2	1
cbronte_jane_part_1	2	1	2	1
cbronte_shirley_part_0	1	1	0	1
cbronte_shirley_part_1	1	1	1	0

Při srovnání textů jednoho autora s texty ostatních autorů je situace obdobná. Počet sdílených typů je velmi nízký, až nulový.

Jako poslední třídu prozkoumáme třídu extrémně nízké frekvence. Tato třída obsahuje velké množství typů s velmi nízkou frekvencí. Disperze slov v textu zde pro určení této

třídy nehraje roli. Sdílení těchto typů mezi autory bude spíše poukazovat na podobnosti v jejich slovní zásobě či podobnost tématu děl.

Uvedeme nejprve matici sdílených typů třídy extrémně nízké frekvence pro jednoho autora.

Tabulka 62: Matice sdílených typů extrémně nízké frekvence u textů jednoho autora

	barclay_ladies_part_0	barclay_ladies_part_1	barclay_rosary_part_0	barclay_rosary_part_1
barclay_ladies_part_0	3798			
barclay_ladies_part_1	1331	3406		
barclay_rosary_part_0	1090	1016	4010	
barclay_rosary_part_1	1105	1045	1351	3745

Na matici můžeme vidět, že texty jednoho autora sdílí okolo jednoho tisíce typů. Opět vidíme více sdílených typů pro texty pocházející ze stejného díla. Texty jednoho autora mezi sebou sdílejí zhruba třetinu typů v třídě extrémně nízké frekvence.

Tabulka 63: Matice sdílených typů extrémně nízké frekvence s texty jiných autorů

	barclay_ladies_part_0	barclay_ladies_part_1	barclay_rosary_part_0	barclay_rosary_part_1
blackmore_erema_part_0	998	843	998	1002
blackmore_erema_part_1	967	844	1018	1007
blackmore_lorna_part_0	949	797	916	878
blackmore_lorna_part_1	1032	881	938	989
braddon_audley_part_0	981	865	1074	1050
braddon_audley_part_1	1004	974	1092	1129
braddon_fortune_part_0	1099	1006	1268	1182
braddon_fortune_part_1	1134	1023	1233	1148
cbronte_jane_part_0	1209	1070	1288	1230
cbronte_jane_part_1	1232	1079	1283	1241
cbronte_shirley_part_0	1226	1100	1288	1280
cbronte_shirley_part_1	1229	1069	1271	1264

Porovnáme-li texty prvního autora s texty ostatních autorů, můžeme si všimnout, že ačkoliv je v některých případech počet sdílených typů obdobný, jako s texty vlastními, může být v některých případech nižší nebo dokonce i značně vyšší než počet sdílených typů mezi texty jednoho autora. Pro extrémně nízce frekventované lexikum tedy

můžeme v některých případech najít větší podobnost mezi texty dvou autorů, než mezi texty jednoho autora.

Shrneme-li tyto poznatky, můžeme dojít k závěru, že jednotlivé třídy se mírně liší v poměru počtu sdílených typů mezi autory a celkovým počtem typů v dané disperzní třídě pro dané texty. Počty sdílených typů mezi texty se také pro jednotlivé třídy dramaticky liší, což je ale dáno především velmi rozdílným celkovým počtem typů v jednotlivých třídách. Zkoumali jsme texty o celkové délce 50000 tokenů dělených na úseky po 5000 tokenech. Pro třídu stabilních jsme pozorovali okolo 200 sdílených typů mezi texty navzájem, což odpovídá zhruba 2/3 typů. U třídy nestabilních jsme pozorovali zhruba 300 – 500 sdílených typů mezi texty. Texty tak mezi sebou navzájem sdílejí zhruba třetinu až polovinu typů ze třídy nestabilní disperze, tedy o něco méně než u třídy stabilní disperze. Ve třídě shluků mezi sebou texty sdílejí pouze jednotky typů, nebo v některých případech nesdílejí vůbec žádné typy. V poměru k počtu typů obsažených ve třídě shluků texty sdílejí přibližně jednu setinu typů klasifikovaných jako disperzní třída shluků. U extrémně nízce frekventovaného lexika zjistíme, že texty mezi sebou sdílejí přibližně 900 až 1500 typů, což odpovídá zhruba 1/4 až 1/3 typů. Největší potenciál k tomu být sdíleny mezi autory mají typy stabilní disperze. Oproti nim mírně nižší potenciál mají typy nestabilní disperze a extrémně nízce frekventovaného lexika, které mají poměr sdílených typů relativně podobný. Minimální potenciál pro sdílení mezi autory mají typy zahrnuté do třídy shluků.

3.7. Využití disperze lexika pro určení autorství

V této části ověříme potenciál slov jednotlivých disperzních tříd pro odlišení autorství textů. Budeme hodnotit jak potenciál disperzních tříd celkově, tak i potenciál užšího výběru lexika z některých disperzních tříd.

Pro nalezení rozdílů mezi dvojicemi textů aplikujeme na každou dvojici textů matematickou funkci, pomocí které kvantifikujeme jejich nepodobnost. Hovoříme-li v tomto kontextu o nepodobnosti, můžeme použít termín vzdálenost jako ekvivalent k termínu nepodobnost. Uvažujeme tedy o nepodobnosti dvou textů jako o vzdálenosti mezi nimi. Pro výpočet vzdálenosti dvou textů lze využít množství různých

matematických funkcí z nichž každá je vhodná pro jiný druh úlohy. V našem případě využijeme tzv. Jaccardovskou vzdálenost.

Jaccardovská vzdálenost měří míru rozdílnosti dvou množin unikátních prvků. Pro texty jde tedy o rozdílnosti dvou množin typů dvou textů. Předpokládejme dva texty p a q. Nechť Sp a Sq značí množiny jejich typů. Jaccardovskou vzdálenost DJ pro Sp a Sq pak vypočteme následovně:

$$DJ\ (Sp, Sq) = \ 1\ - \frac{|Sp\ \cap\ Sq|}{|Sp\ \cup\ Sq|}$$

$|Sp\ \cap\ Sq|$ odpovídá průniku obou množin. Tedy množině typů sdílených oba texty. $|Sp\ \cup\ Sq|$ odpovídá sjednocení obou množin. Ted množině všech typů vyskytujících se v jednom z textů či v obou textech. Vzdálenost může nabývat hodnot mezi 0 – 1, přičemž nulová vzdálenost značí úplnou shodu mezi dvěma texty, zatímco vzdálenost dosahující hodnoty 1 znamená žádnou shodu mezi porovnávanými texty. (Shade & Altmann, 2023)

Potenciálními nevýhodami Jaccardovské vzdálenosti je, že nezohledňuje počet výskytů slov. Pracuje pouze s unikátními slovy a měří tedy pouze podobnost slovníku dvou textů. Další nevýhodou je vyšší citlivost na rozdíly v délce porovnávaných textů. (Shade & Altmann, 2023)

V případě naší práce budeme vypočítávat Jaccardovskou vzdálenost nikoliv pro všechny typy daných textů, ale pro omezené výběry typů splňujících dané podmínky, jakými budou například příslušnost ke určité disperzní třídě, maximální či minimální frekvence výskytů apod. Zohledníme zde pouze výskyt typu. Necitlivost Jaccardovské vzdálenosti k frekvenci výskytů pro nás tedy nebude překážkou. Zároveň, jelikož jsou všechny porovnávané texty upraveny na totožnou vzdálenost, ani vyšší citlivost k rozdílům v délce textu zde nebude problémem.

Pro zhodnocení potenciálu k určení autorství textu budeme postupovat následovně. Nejprve si zvolíme výběr textů, jenž obsahuje texty více autorů, z nichž každý je reprezentován několika texty. Pro texty získáme seznamy jejich typů spolu s informacemi o frekvenci jejich výskytů, příslušnosti k disperzní třídě. Na tyto seznamy typů aplikujeme zvolený filtr. Kupříkladu si můžeme ponechat pouze typy určité disperzní třídy nebo dosahující určitého rozpětí frekvencí výskytů. Pro každý text

nám tak zůstane seznam typů, který splňuje podmínky k aplikaci zvoleného filtru. Pro každý text reprezentovaný filtrovaným seznamem typů vypočteme Jaccardovskou vzdálenost mezi ním a každým jiným textem z výběru, který je rovněž reprezentován filtrovaným seznamem typů. Výsledkem tohoto procesu bude matice vzdáleností pro každý text z výběru s každým jiným textem z výběru. Matice bude obsahovat hodnoty od 0 do 1, přičemž nulová vzdálenost značí, že jsou seznamy typů textů naprosto shodné, zatímco vzdálenost 1 znamená, že jsou seznamy typů textů maximálně odlišné a žádné typy mezi sebou navzájem nesdílejí. Pomocí matice vzdáleností můžeme nepodobnost seznamů typů, které texty reprezentují, vizualizovat využitím metod, které popíšeme níže. Skrz volbu různých filtrů pro seznamy typů a vizualizaci nepodobností mezi texty můžeme posoudit, které druhy lexika umožní efektivněji odlišit různé autory.

Jednou z metod, kterou aplikujeme na matice vzdáleností, bude tzv. vícerozměrné škálování. Vícerozměrné škálování umožňuje umístit data, v našem případě texty reprezentované seznamy typů, jako body v dvojrozměrném prostoru. Výsledná vizualizace tak připomíná mapu, na které jsou texty zobrazeny jako body. Texty blízko sebe jsou sobě navzájem podobnější než texty od sebe výrazněji vzdálené. Pokud jsou texty napsané stejným autorem více podobné sobě navzájem, potom očekáváme, že budou vícerozměrným škálováním zobrazeny jako skupina blízkých bodů. Nevýhodou tohoto přístupu je, že zobrazení na dvojrozměrné ploše neumožňuje zachytit variaci v původních datech s perfektní přesností. Použitím vícerozměrného škálování vždy ztratíme určité množství informace obsažené v datech. Přestože se této ztrátě informací u vícerozměrného škálování vyhnout nemůžeme, máme možnost změřit, do jaké míry vystihuje naše zobrazení původní data. K tomuto účelu využijeme metriku nazývanou „goodness of fit“. Nízké hodnoty budou znamenat, že body zobrazené blízko sebe se vyznačují značnou podobností i v původních datech matice vzdáleností. Naopak vysoké hodnoty této metriky znamenají, že blízké body nutně nemusejí být podobné a zobrazení je proto nespolehlivé. Obecně tedy platí, že čím je hodnota fitu bližší nule, tím lépe a spolehlivěji vystihuje původní data. Vícerozměrné škálování nám nepomůže definitivně určit nejpodobnější texty, ale umožní nám vizualizovat rozdíly mezi texty. Analýza využívající vícerozměrné škálování tedy bude mít převážně explorativní charakter. (Davidson & Sireci, 2000)

Druhou metodou, pomocí které vizualizujeme hodnoty v maticích vzdáleností, bude tzv. hierarchické shlukování. Tato metoda nám umožňuje dát odpověď na otázku, které texty jsou si navzájem nejblíže. Pomocí hierarchického shlukování vytvoříme pro matici vzdáleností binární strom. Hierarchické shlukování postupuje od jednotlivých datových elementů, přičemž vždy vyhledá dva datové elementy, které jsou navzájem nejblíže, a následně je sloučí k sobě. Takto postupuje dále, dokud nejsou všechny datové elementy navzájem propojeny. Grafickou reprezentaci výsledného binárního stromu nazýváme dendrogram. Na výsledném dendrogramu pak můžeme jasně vidět, které datové elementy, kterými jsou v našem případě texty, jsou sobě navzájem nejblíže. K určení vzdálenosti mezi shluky datových elementů lze využít více různých metod. Základní běžně používané metody určení vzdálenosti, anglicky nazývané „linkage methods“, jsou tři. První z nich je metoda nejbližšího souseda, která používá vždy vzdálenost od nejbližšího datového elementu ve shluku. Její nevýhodou je, že může dojít k řetězení elementů, což může být nevýhodné pro identifikaci skupin. Druhou metodou je metoda nejvzdálenějšího souseda, která za vzdálenost shluku považuje vždy nejvíce vzdálený datový element ve shluku. Tato metoda většinou vede k výraznějšímu oddělení shluků od sebe a předchází řetězení. Třetí metodou je metoda průměrné vzdálenosti, kdy je za vzdálenost od shluku považována průměrná vzdálenost všech datových elementů ve shluku. Nevýhodou průměrné metody je, že v případě, kdy jsou mezi jednotlivými datovými elementy velké rozdíly, nemusí být průměrná vzdálenost odpovídající pro většinu z nich. Jelikož se v rámci našeho výzkumu pokoušíme co nejlépe oddělit texty jednoho autora od ostatních, použijeme metodu nejvzdálenějšího souseda, která je anglických termínem nazývána jako „complete linkage. (Nielsen, 2016)

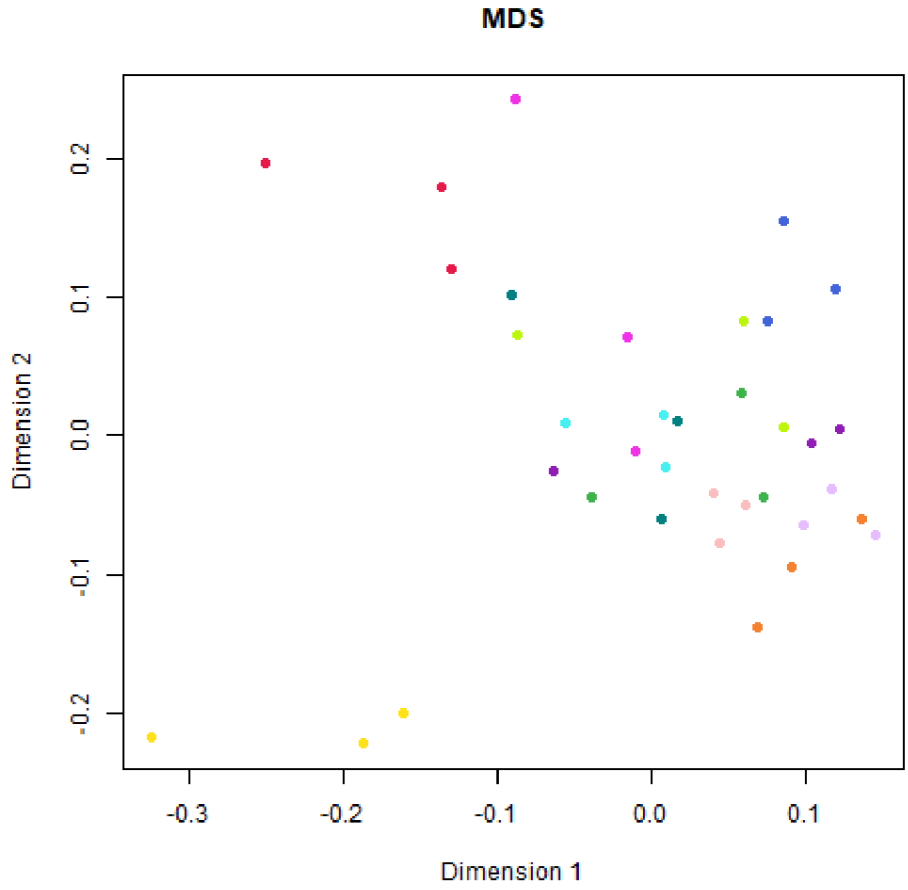
Pro potřeby určování autorství si vybereme širší výběr textů s více autory, než který jsme využívali doposud. Tento výběr bude zahrnovat dvanáct různých autorů, z nichž každý bude reprezentován třemi texty pocházejícími vždy z odlišného díla. Pro tento výběr půjde o texty dlouhé 50000 tokenů dělené na úseky po 5000 tokenech.

Nejprve použijeme pro určení autorství textů výběry typů omezené na jednotlivé disperzní třídy. Kromě příslušnosti k disperzní třídě prozatím nepoužijeme jiná kritéria pro filtrování typů. Abychom však mohli zhodnotit vlastnosti disperzních tříd pro určení autorství, musíme nejprve stanovit základní metodu, vůči které vlastnosti použitých filtrů porovnáme. Touto metodou bude vícerozměrné škálování a

hierarchické shlukování použité na matici vzdáleností vytvořené na základě všech typů obsažených v textu bez užití filtrů. Naším cílem zde bude najít takový druh filtru, který nám umožní maximálně efektivně odlišit autory stejných textů od sebe použitím co nejmenšího množství typů. Nejprve tedy uvedeme bodový graf vícerozměrného škálování a dendrogram zobrazující hierarchické shlukování pro všechny typy textů bez aplikování filtru.

Níže je uveden bodový graf vícerozměrného škálování zkonstruovaný použitím všech typů bez aplikace filtrů. Průměrný počet typů v textu je tak 5627 typů. Texty pocházející od stejného autora jsou v bodovém grafu vizualizovány pomocí stejné barvy.

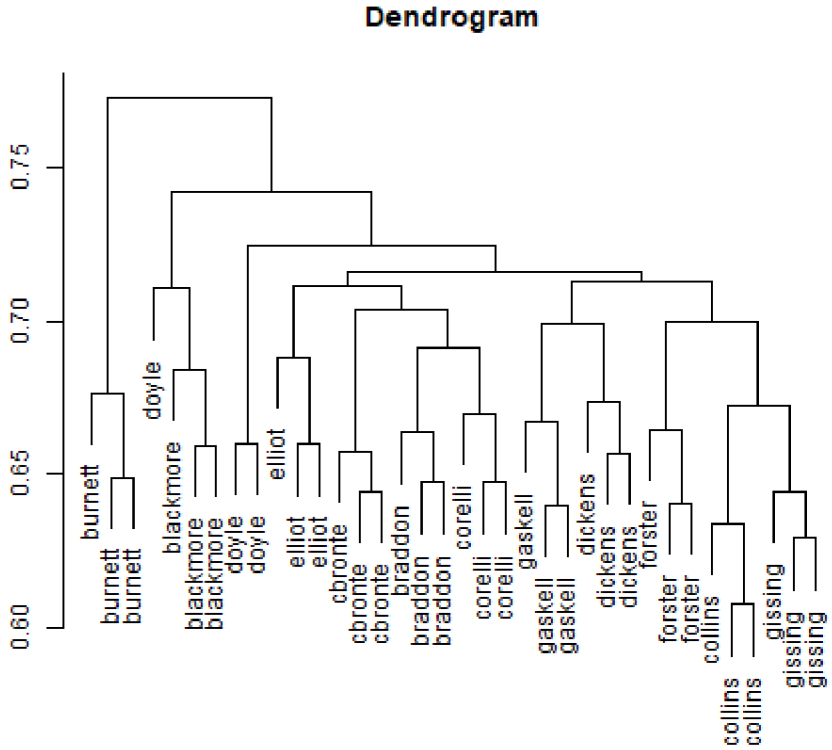
Obrázek 1: Vícerozměrné škálování pro všechny typy



Na bodovém grafu vícerozměrného škálování vidíme, že skupiny textů náležejících jednomu autorovi se do značné míry překrývají. Texty stejného autora jsou však ve většině případů umístěny poměrně blízko sebe. Vizuální odlišení skupin textů náležících jednomu autorovi je obtížné. Metrika „goodness of fit“ dosahuje hodnoty 0.101851.

Níže je uveden dendrogram zobrazující hierarchické shlukování pro stejná data.

Obrázek 2: Dendrogram pro všechny typy

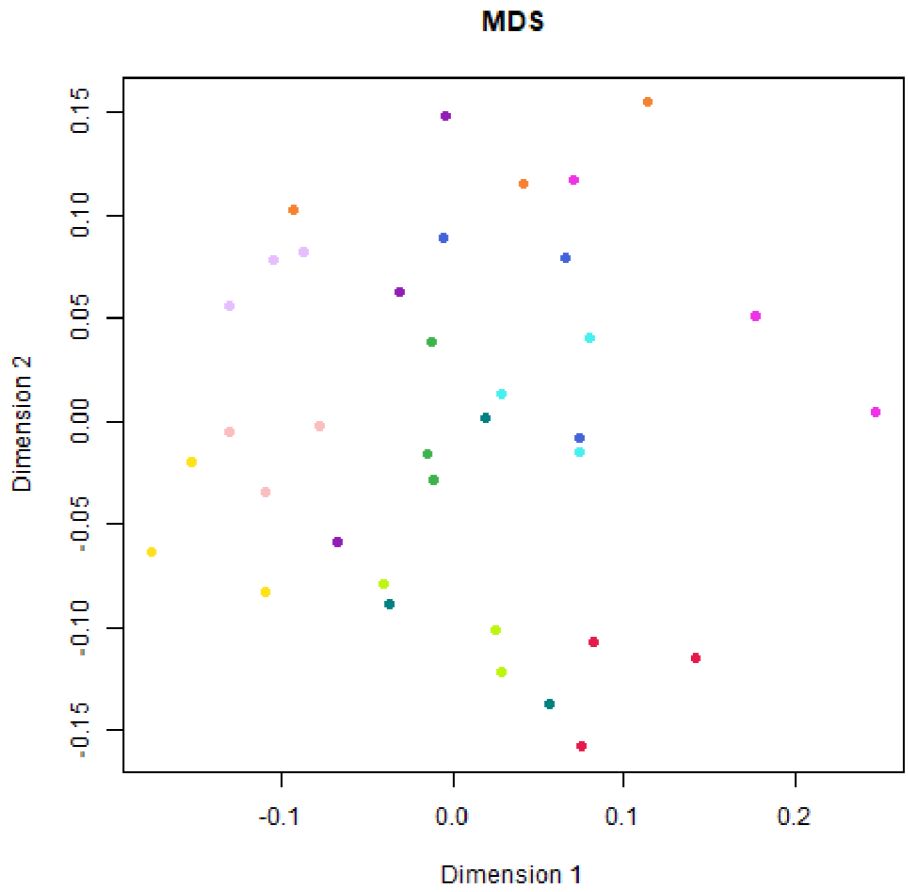


Na základě hierarchického shlukování se v tomto případě povedlo určit všechny tři texty jednoho autora jako nejbližší texty pro 11 z 12 autorů s využitím průměrně 5627 typů na text. Bez využití filtrů, tedy pomocí Jaccardovské vzdálenosti, nalezneme texty sdílející autora relativně úspěšně. Měříme zde však spíše podobnost slovníků užívaných v textech autorů. Abychom nalezli lexikum nesoucí autorskou informaci, musíme počet tokenů užitý k určení autorství snížit. Následně se tedy pokusíme pomocí filtrování typů zredukovat počet typů využívaných k určení autorství tak, abychom zachovali co nejlepší schopnost přiřazení textů stejného autora k sobě navzájem.

Nyní se zaměříme na disperzní třídu lexika se stabilní disperzí. Pro každý text ve výběru získáme pomocí filtru pouze typy se stabilní disperzí. Průměrný počet typů na text tak dosahuje hodnoty 280 typů.

Níže je zobrazen bodový graf vícerozměrného škálování pro typy stabilního lexika.

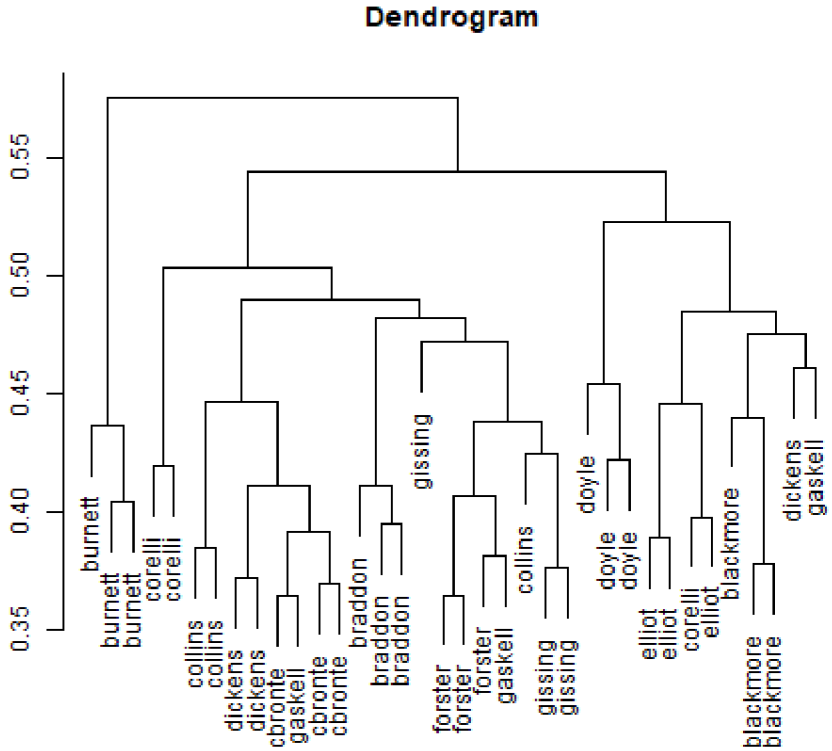
Obrázek 3: Vícerozměrné škálování pro všechny typy stabilní disperze



V grafu si můžeme povšimnou značného překryvu mezi texty různých autorů. Pro některé autory platí, že jsou jejich texty zobrazeny relativně blízko sebe, ale v řadě případů se do velké míry překrývají i s texty jiných autorů. Zobrazení nám neumožňuje snadné vizuální odlišení autorů. Hodnota metriky „goodness of fit“ pro tento graf dosahuje 0.1568615.

Níže je uveden dendrogram zobrazující hierarchické shlukování pro stejná data.

Obrázek 4: Dendrogram pro všechny typy stabilní disperze

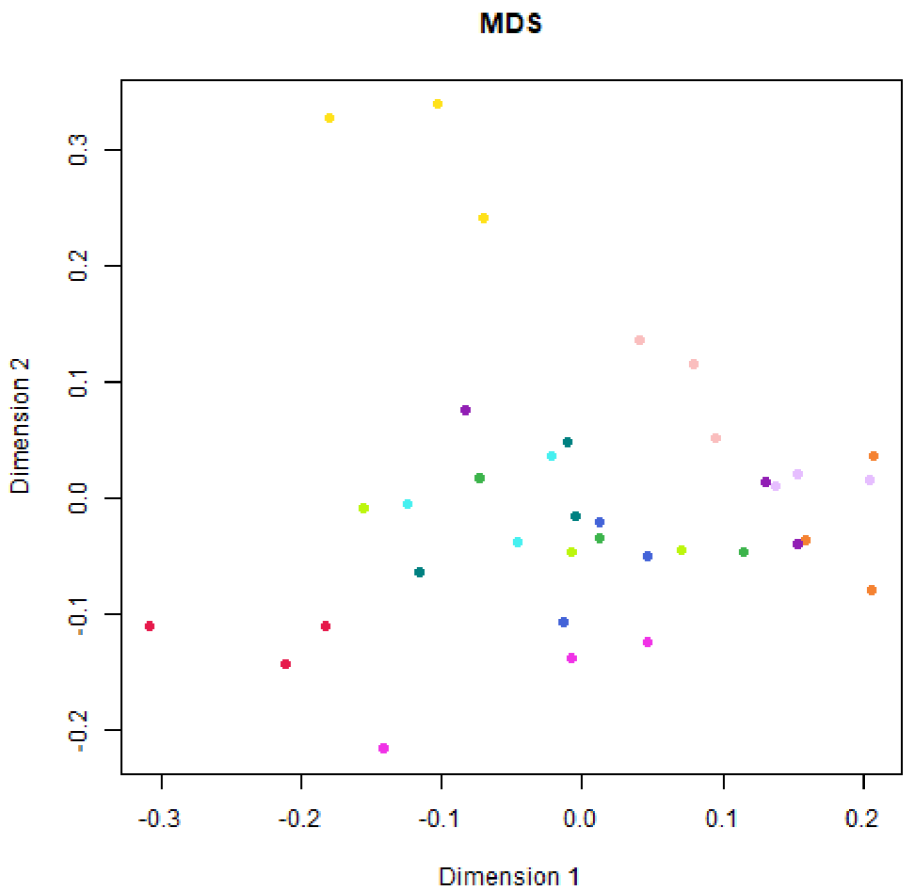


Pomocí tohoto dendrogramu vytvořeného na základě lexika se stabilní disperzní přiřadíme texty stejného autora k sobě jen v části případů. Všechny tři texty jednoho autora se zde povedlo určit jako nejbližší sousedy pro 6 z 12 autorů s využitím průměrně 280 typů na text. Pro většinu autorů se povedlo určit alespoň dva texty stejného autora jako nejbližší sousedy.

Dále se podíváme na vlastnosti lexika třídy nestabilní disperze. Analogicky s předchozím postupem použijeme k filtrování pouze příslušnost k disperzní třídě. Po aplikování filtru bude průměrný počet typů na text dosahovat hodnoty 1024 typů.

Níže uvedeme bodový graf vícerozměrného škálování s použitím typů náležejících do třídy nestabilní disperze.

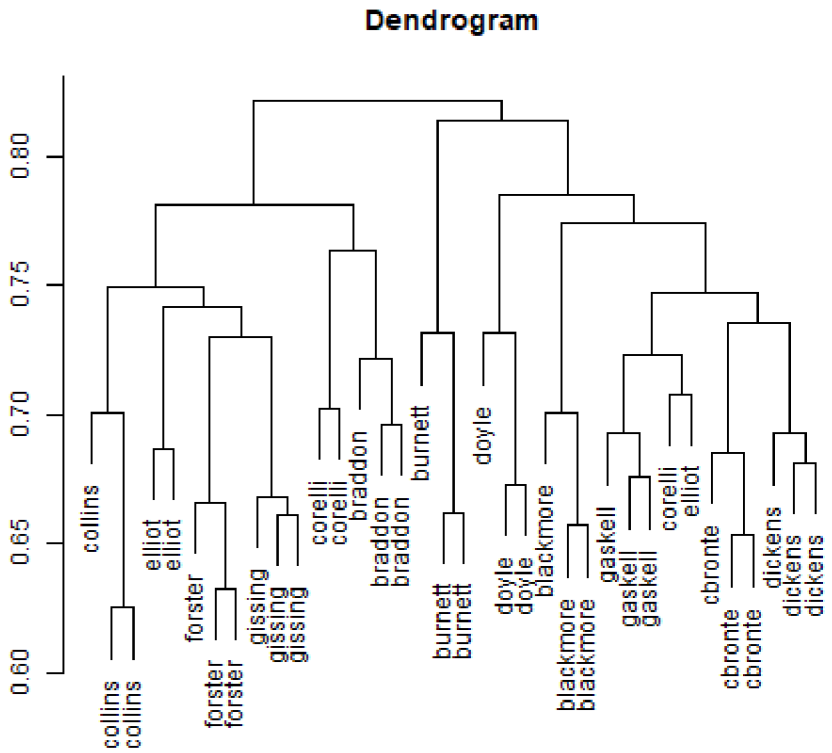
Obrázek 5: Vícerozměrné škálování pro všechny typy nestabilní disperze



U typů třídy nestabilní disperze vidíme značný překryv mezi jednotlivými autory, obdobně jako u typů stabilní disperze. Můžeme si povšimnout, že texty jsou méně rozprostřené než v předchozím případě. Zobrazení nám neumožňuje snadné vizuální odlišení autorů. Metrika „goodness of fit“ dosahuje hodnoty 0.1120545. Vizualizace tedy vystihuje původní data o něco lépe, než u stabilní disperze. U nestabilní disperze máme k dispozici více jak trojnásobné množství typů.

Níže je uveden dendrogram zobrazující hierarchické shlukování pro vzdálenosti textů na základě typů třídy nestabilní disperze.

Obrázek 6: Dendrogram pro všechny typy nestabilní disperze

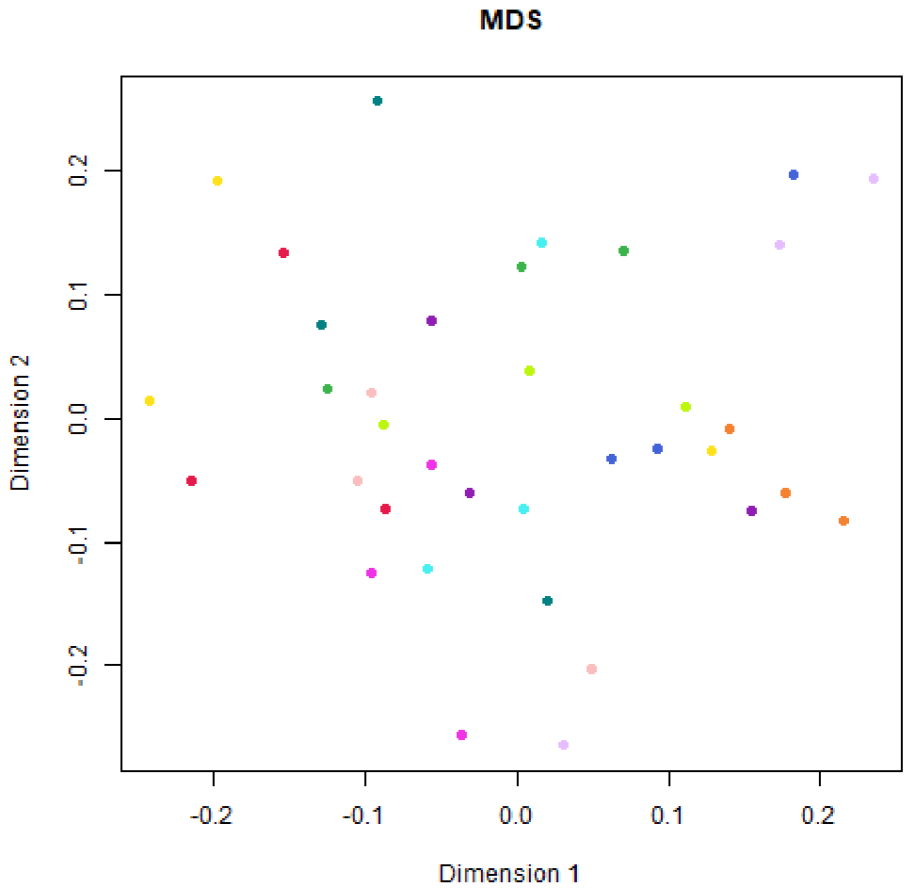


Na základě hierarchického shlukování se v tomto případě povedlo určit všechny tři texty jednoho autora jako nejbližší texty pro 10 z 12 autorů s využitím průměrně 1024 typů na text. Pomocí typů s nestabilní disperzí lze tedy sdružit texty autorů relativně úspěšně, avšak počet využitých typů je poměrně vysoký.

Následující disperzní třídou, na kterou se zaměříme, je třída shluků. Dříve jsme již uvedli, že u třídy shluků je jen velmi malý počet sdílených typů mezi texty. Průměrný počet typů na text dosahuje hodnoty 100 typů.

Níže je zobrazen bodový graf vícerozměrného škálování pro typy třídy shluků.

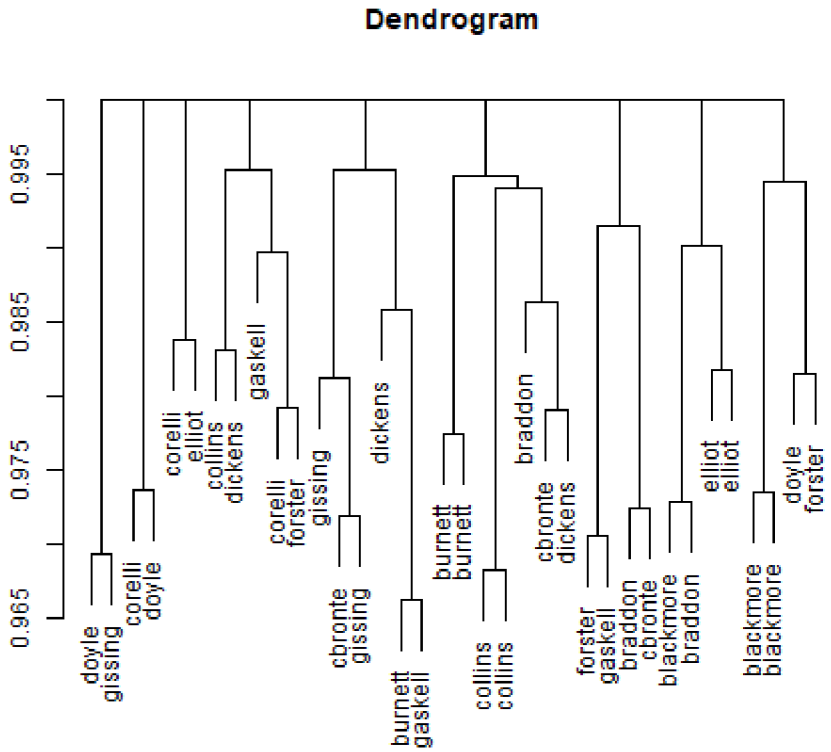
Obrázek 7: Vícerozměrné škálování pro všechny typy třídy shluků



Pro třídu shluků jsou texty v bodovém grafu vícerozměrného škálování rozloženy způsobem, který neumožňuje identifikovat skupiny příslušející textům jednotlivých autorů. Lexikum třídy shluků se jeví jako nevhodné pro využití vícerozměrného škálování k identifikaci autorství textů. Metrika „goodness of fit“ dosahuje hodnoty 0.06453756.

Níže je uveden dendrogram zobrazující hierarchické shlukování pro vzdálenosti textů na základě typů třídy shluků.

Obrázek 8: Dendrogram pro všechny typy třídy shluků

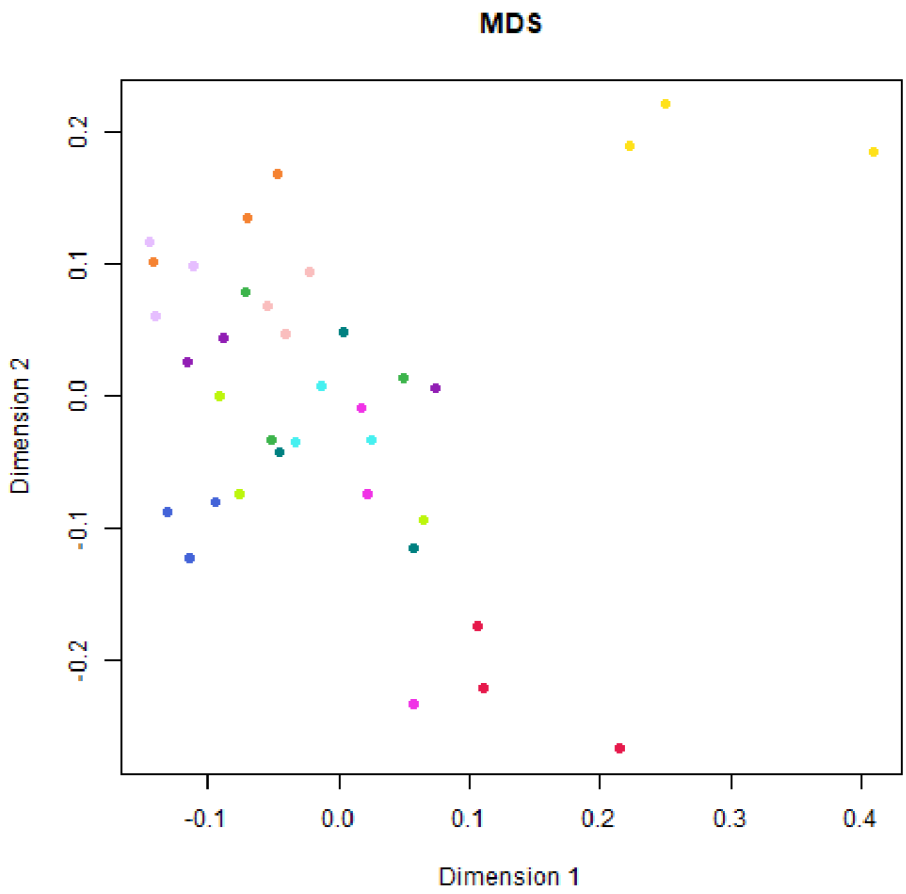


Hierarchické shlukování pro texty reprezentované lexikem třídy shluků nedokáže k sobě spolehlivě přiřadit texty náležející stejným autorům. Všechny tři texty jednoho autora se zde povedlo určit jako nejbližší sousedy pro 0 z 12 autorů s využitím průměrně 100 typů na text. Třída shluků se tedy jeví jako nevhodná k určování autorství.

Dále se zaměříme na třídu extrémně nízkofrekventovaného lexika. Jelikož se jedná o slova s počtem výskytů tři a méně, představuje tato třída většinu typů textu. Výsledky vícerozměrného škálování a hierarchického shlukování tak budou vypovídat spíše o podobnostech slovníků autorů. Průměrný počet typů na text je 4222 typů.

Níže uvedeme bodový graf vícerozměrného škálování s použitím typů náležících do třídy extrémně nízkofrekventovaného lexika.

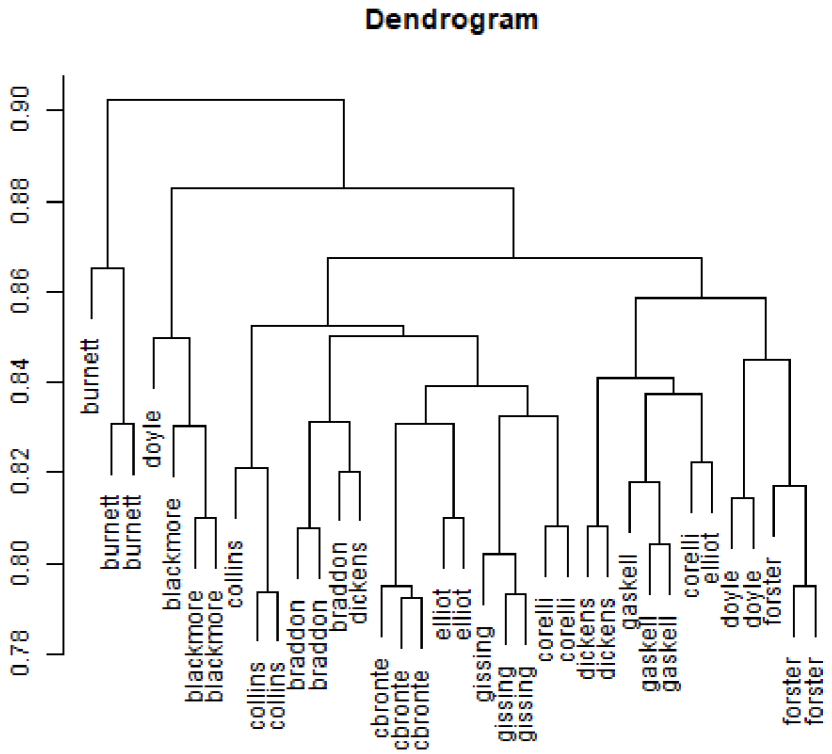
Obrázek 9: Vícerozměrné škálování pro všechny typy třídy extrémně nízkofrekventovaného lexika



U bodového grafu vícerozměrného škálování pomocí typů třídy extrémně nízkofrekventovaného lexika můžeme vidět, že většina textů jednoho autora je zobrazena poměrně blízko sebe. Zároveň však dochází ke značnému překryvu mezi texty různých autorů. Vizuálně je tedy stále obtížné odlišit texty různých autorů od sebe. Metrika „goodness of fit“ dosahuje hodnoty 0.08337214.

Níže je uveden dendrogram zobrazující hierarchické shlukování pro vzdálenosti textů na základě typů třídy extrémně nízkofrekventovaného lexika.

Obrázek 10: Dendrogram pro všechny typy třídy extrémně nízkofrekventovaného lexika



Hierarchické shlukování pro texty reprezentované lexikem třídy extrémně nízkofrekventovaného lexika sdružuje texty stejného autora ve většině případů. Všechny tři texty jednoho autora se zde povedlo určit jako nejbližší sousedy pro 8 z 12 autorů s využitím průměrně 4222 typů na text. Odlišení autorů od sebe není spolehlivé. Počet využitých typů je velmi vysoký. Rozdíly mezi autory jsou pravděpodobně dány spíše rozdíly ve slovní zásobě užívané v textech než přítomností pro autora specifického lexika.

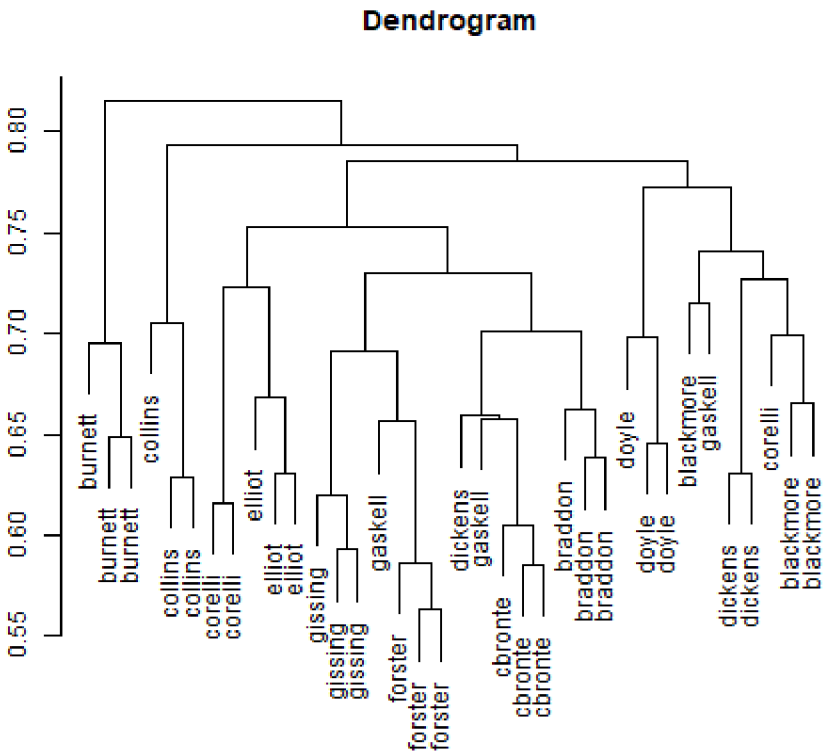
Shrneme-li dosavadní výsledky, zjistíme o vlastnostech posuzovaného lexika následující. Na základě vícerozměrného škálování se nám nepodařilo vizuálně oddělit díla jednotlivých autorů od sebe, aniž by se navzájem překrývala. Hierarchické shlukování oproti tomu poskytuje snáze hodnotitelné výsledky. Nejlepší výsledky jsme zaznamenali u lexika třídy nestabilní disperze. Pomocí ní jsme určili všechny tři texty jednoho autora jako nejbližší pro 10 z 12 autorů, přičemž jsme brali v potaz průměrně 1024 typů na text. V porovnání s nevyužitím žádných filtrů jde však stále o horší výsledek. Hierarchické shlukování bez využití filtrů určilo díla stejného autora jako nejbližší pro 11 z 12 autorů s využitím průměrně 5627 typů na text. Třída nestabilní však poskytla podobné výsledky s využitím výrazně menšího množství typů. Pomocí typů třídy extrémně nízkofrekventovaného lexika jsme určili díla stejného autora jako nejbližší pro 8 z 12 autorů s použitím průměrně 4222 typů. Jde tedy o horší výsledek než u nefiltrovaných typů, přestože byl počet užitých typů zredukován poměrně málo. Pomocí třídy lexika se stabilní disperzí jsme určili díla stejného autora jako nejbližší pro 6 z 12 autorů s použitím průměrně 280 typů na text. Úspěšnost tedy byla poměrně nízká, nicméně v poměru k počtu užitých typů bychom mohli očekávat úspěšnost mnohem nižší. Domníváme se tedy, že by toto lexikum mohlo nést hodnotnou autorskou informaci. Nejhorší výsledky měla třída shluků. Pomocí lexika třídy shluků jsme určili díla stejného autora jako nejbližší pro 0 z 12 autorů s použitím průměrně 100 typů na text. Tato třída se zdá být nevhodnou pro určení autora textu.

Nejlepší potenciál pro určování autorství textu tedy nacházíme u třídy lexika s nestabilní disperzí a třídy lexika se stabilní disperzí, jelikož vykazují nejlepší poměr úspěšnosti identifikace skupin textů jednoho autora a počtu typů, které jsou k tomu využity. Dále se tedy zaměříme právě na tyto dvě disperzní třídy.

Pro bližší analýzu vlastností lexika tříd stabilní a nestabilní disperze se omezíme na provedení hierarchického shlukování. Budeme postupovat přidáváním dalších restrikcí pro filtrování typů.

Nejprve se zaměříme na třídu lexika se stabilní disperzí. Získáme pouze typy třídy stabilní disperze s celkovou frekvencí výskytů nižší než 60 výskytů. Takto získáme přibližně typy s frekvencí blízkou mediánu a nižší.

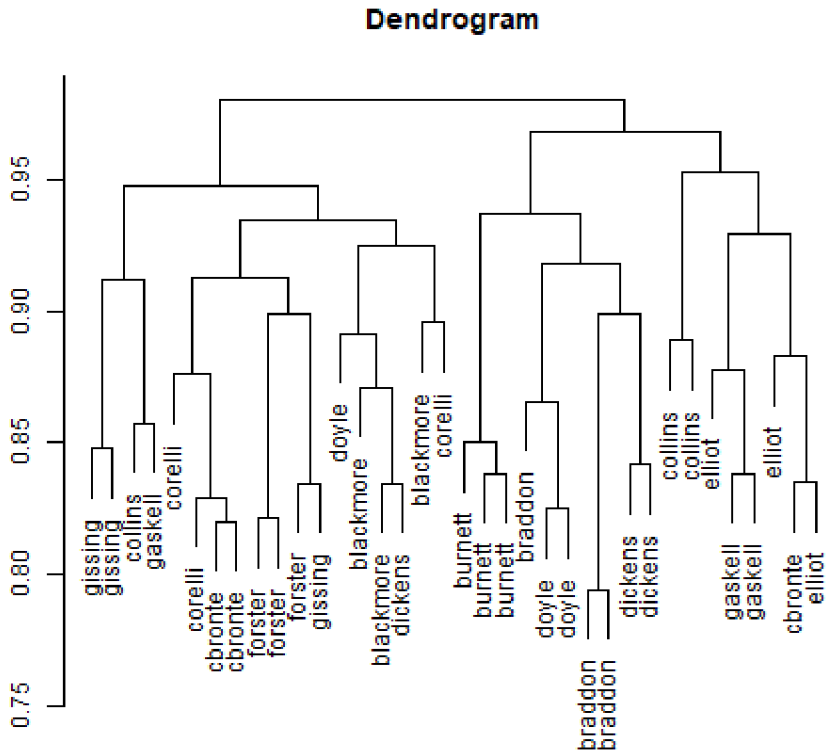
Obrázek 11: Dendrogram pro lexikum stabilní disperze s hranicí 60 výskytů



Správně tak určíme texty pro 8 z 12 autorů. Využijeme k tomu průměrně 166 typů. Omezením výběru frekvenčním limitem v rámci třídy stabilní disperze jsme určili texty přesněji než bez frekvenčního limitu, kdy jsme určili správně texty pouze pro 6 z 12 autorů. Se snížením množství typů z původních průměrně 280 typů na 166 typů došlo ke zlepšení úspěšnosti určení autorství.

Následovně snížíme hranici maximálního počtu výskytů pro lexikum třídy stabilní disperze na 30 výskytů.

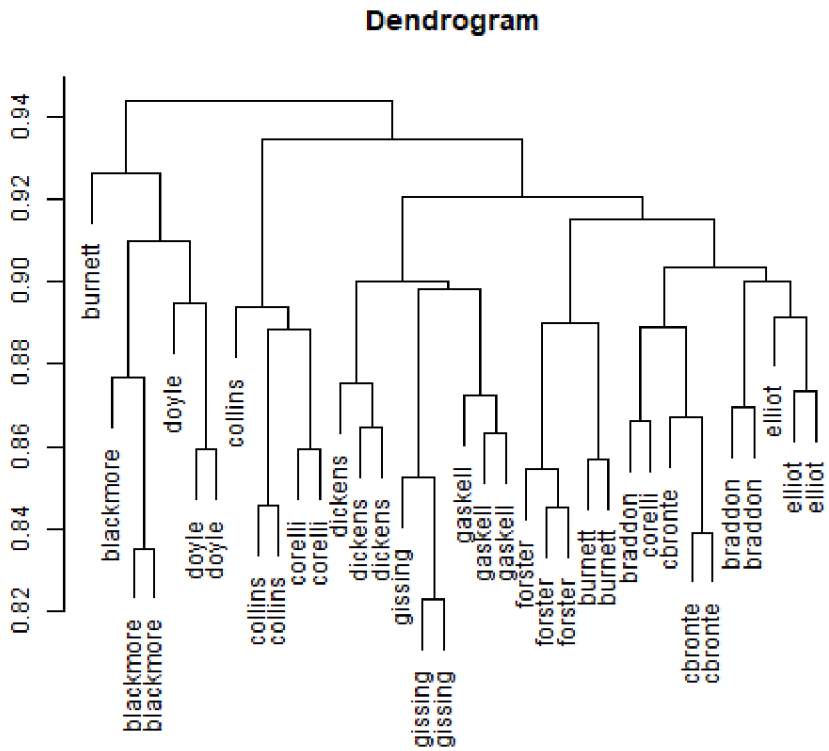
Obrázek 12: Dendrogram pro lexikum stabilní disperze s hranicí 30 výskytů



V tomto případě jsme správně určili všechny tři texty jednoho autora jako nejbližší pouze pro 1 autora z 12 autorů. Využili jsme průměrně 77 typů na text. Další snižování tedy nevykazuje zlepšení jako v předchozím případě.

Nyní se zaměříme na lexikum s nestabilní disperzí. Analogicky s předchozím postupem vybereme pouze typy lexika se stabilní disperzí v oblasti mediánu a nižší. Zachováme tedy pouze typy s frekvencí 8 výskytů a nižší frekvencí.

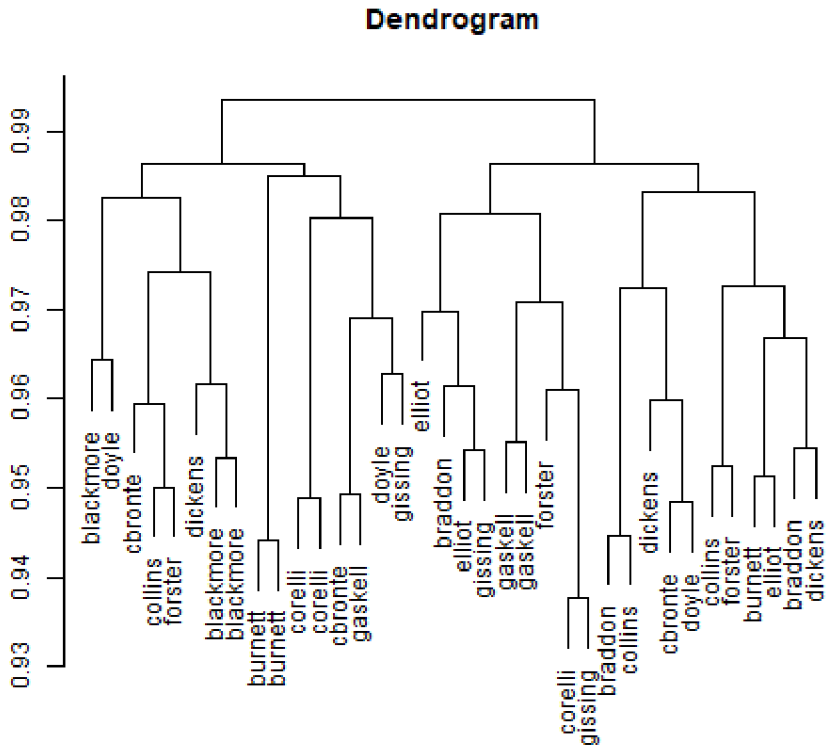
Obrázek 13: : Dendrogram pro lexikum nestabilní disperze s hranicí 8 výskytů



Na základě tohoto výběru určíme úspěšně všechny tři texty jednoho autora jako nejbližší pro 9 autorů ze 12 autorů, přičemž využijeme průměrně 587 typů na text. Bez zavedení horního maxima frekvencí jsme pro třídu nestabilní frekvence správně určili texty pro 10 z 12 autorů s využitím průměrně 1024 typů na text. Použitím limitu frekvence výskytů jsme tedy výrazně snížili počet použitých typů, přičemž se přesnost určení autorství se mírně snížila. Tento rozdíl však mohl být způsoben i vlivem náhody.

Následovně analogicky s přechozím postupem dále snížíme limit frekvence výskytů pro třídu nestabilní disperze. Zvolíme limit frekvence 5 výskytů.

Obrázek 14: Dendrogram pro lexikum nestabilní disperze s hranicí 5 výskytů



Na základě tohoto výběru určíme úspěšně všechny tři texty jednoho autora jako nejbližší pro 0 autorů ze 12 autorů, přičemž využijeme průměrně 212 typů na text. Snižování limitu maximální frekvence výskytů do této míry se ukazuje jako detrimentální pro úspěšné klastrování textů stejných autorů.

4. Závěr

Nyní v závěru této diplomové práce shrneme poznatky našeho výzkumu.

Naším cílem bylo studovat, jaké vlastnosti má u lexika jeho disperze v rámci textu a pokusit se využít disperzi lexika jako stylometrickou metriku pro určování autorství textů. K měření disperze lexika jsme využili metodu spočívající v rozdělení textu na více stejně dlouhých úseků a měření výskytu frekvence v jednotlivých úsecích. Jako nejvíce užitečnou měřitelnou vlastnost považujeme počet prázdných úseků, tedy počet úseků s nulovou frekvencí výskytů. Obdobně lze použít i počet úseků s nenulovým počtem výskytů. Doplníme-li informaci o celkovém počtu výskytů slova informací o počtu úseků, ve kterých se vyskytuje či nevyskytuje, získáme jasnější představu o roli lexika v textu. Převážně na základě počtu prázdných úseků jsme dále rozdělili lexikum do čtyř různých tříd, které zde nazýváme jako disperzní třídy.

Třída extrémně nízkofrekventovaného lexika je definována pouze na základě celkového počtu výskytů. Představuje lexikum příliš málo frekventované na to, aby měla jeho klasifikace pomocí počtu obsazených úseků smysl. Zahrnuje tedy převážně pravé hapax legomena, dis legomena a tris legomena. Třída zahrnuje přibližně 80 % všech typů, ale jen zhruba 15 % tokenů. Nejběžnější jsou hapax legomena, tedy slova s právě jedním výskytem.

Třída shluků je lexikum koncentrující se v rámci jednoho úseku či na hranici dvou úseků. Zahrnuje přibližně 2 % všech typů v textu a zhruba 1 % všech tokenů. Jedná se o velmi málo zastoupenou třídu. Nejběžnější jsou shluky s frekvencí čtyř výskytů. Většinou zde převažují substantiva.

Třída stabilní disperze obsahuje lexikum, které se objevuje v naprosté většině úseků. Zahrnuje vysoce frekventovaná funkční slova i slova obsahová. Vzhledem k tomu, že nejvíce frekventovaná slova spadají do této třídy, má tato třída velký počet tokenů vůči typům. Zahrnuje přibližně 5 % všech typů a zhruba 60 % tokenů. Ve vysokých frekvencích výskytu převažují funkční slova. Nejběžnější je lexikum s frekvencí 40 – 55 výskytů. V těchto frekvencích nacházíme lexikum různých slovních druhů, a to lexikum obsahové i funkční.

Třída nestabilní disperze představuje přechodnou třídu mezi třídou stabilní disperze a třídou shluků. Lexikum s tímto typem disperze najdeme ve větším množství úseků, ale v alespoň několika úsecích se nevyskytuje. Třída zahrnuje přibližně 15 % všech typů a zhruba 20 % všech tokenů v textu. Nejběžnější je lexikum s frekvencí 6 – 7 výskytů. Jde převážně o substantiva a verba.

Pro zhodnocení toho, do jaké míry je lexikum v některé z disperzních tříd unikátní pro určitého autora, jsme použili metodu hierarchického shlukování a metodu vícerozměrného škálování pro matice nepodobnosti vypočítané pomocí Jaccardovské vzdálenosti pro různé výběry lexika založené na disperzních třídách a celkové frekvenci. Tyto metody jsme aplikovali na výběr 36 textů zahrnujících 12 různých autorů. K porovnání jsme tyto metody analogicky aplikovali i na všechny typy bez využití filtrů. Na základě všech typů jsme skrze hierarchické shlukování s použitím tzv. „complete linkage“ metody určili texty stejných autorů jako nejbližší poměrně spolehlivě. Správně se nám povedlo určit všechny tři texty jednoho autora jako nejbližší texty pro 11 z 12 autorů s využitím průměrně 5627 typů na text. Cílem tedy bylo získat podobně dobré výsledky s užitím co nejmenšího množství typů. Metoda vícerozměrného škálování nedokázala jasně oddělit skupiny textů jednoho autora od sebe pro žádný z výběrů lexika. Vzájemné vzdálenosti zobrazení bodů se pro různé výběry lexika lišily, nicméně vždy docházelo k překryvu bodů reprezentujících texty různých autorů. Metoda vícerozměrného škálování nám tedy neumožnila vizuálně snadno identifikovat texty jednoho autora.

Užití lexika třídy extrémně nízkofrekventovaného lexika mělo za následek zhoršení klastrování oproti nefiltrovanému výběru, přičemž počet využitých typů zůstal stále velmi vysoký.

Užití lexika třídy shluků neumožnilo určit správně texty jednoho autora ani pro jednoho z nich. Třída shluků se tak pro určování autorství jeví jako absolutně nevhodná.

Užití lexika třídy stabilní disperze umožnilo správně určit texty jednoho autora pro 6 z 12 autorů, s využitím průměrně 280 typů na text. Dalším omezením tohoto výběru pomocí limitu frekvence zhruba v oblasti mediánu pro třídu stabilní disperze jsme dosáhli zlepšení klastrování. Užitím výběru lexika třídy stabilní disperze s celkovou frekvencí nižší než 60 výskytů jsme určili texty jednoho autora správně pro 8 z 12 autorů s využitím průměrně 166 typů. Dalším snižováním maximální frekvence

výskytů se již výsledky klastrování nezlepšily. Tyto výsledky však naznačují, že vysoce frekventovaná funkční slova obsažená v této třídě nebyla rozhodujícím faktorem pro správné klastrování textů autorů, jelikož jejich odstraněním se výsledky klastrování mírně zlepšily.

Užití lexika nestabilní disperze umožnilo správně určit texty jednoho autora pro 10 z 12 autorů s využitím průměrně 1024 typů na text. Vykazuje tedy jen mírné zhoršení přesnosti klastrování, avšak využívá výrazně menší množství typů k dosažení srovnatelného výsledku. Dalším omezením výběru zavedením maximálního počtu výskytů přibližně v oblasti mediánu se přesnost klastrování mírně snížila. Správně byly určeny texty jednoho autora pro 9 z 12 autorů s využitím průměrně 587 typů na text. Přestože se přesnost klastrování mírně zhoršila, opět se povedlo značně zredukovat počet užitých typů. Dalším snižováním maximálního počtu výskytů se výsledky pouze zhoršily.

Lexikum se stabilní disperzí, a lexikum s nestabilní disperzí se tedy jeví jako nejvíce výhodné pro potřeby určování autorství. Jejich užití nám neumožnilo zlepšit přesnost klastrování oproti výsledkům výběru zahrnujícího všechny typy v textu, nicméně prokazuje, že typy stabilní a nestabilní disperze jsou více distinktivní pro styl autorů než lexikum jiných tříd. Umožňuje nám to významně omezit výběr typů, který analyzujeme. Disperze lexika se jako samostatná metrika neukazuje jako příliš efektivní pro určování autorství. Kombinací metrik disperze a frekvence výskytů lze oproti tomu dosáhnout výrazně lepších výsledků. Potenciál kvantitativního měření disperze tak vidíme zejména v doplnění jiných metrik. Domníváme se tedy, že naše výsledky podporují naši výchozí hypotézu, že studium vlastností disperze daného typu uvnitř daného textu nám může poskytnout informace nad rámce poznatků, které bychom získali studiem pouze frekvence výskytů. Vzájemný vztah mezi disperzí a frekvencí výskytů je však značně silný a úplné oddělení těchto vlastností není žádoucí. Náš výzkum ukazuje, že průběžně opakující se lexikum s nízkou frekvencí je více distinktivní pro autorský styl, než jiné lexikum. V tomto se naše poznatky shodují se závěry výzkumu Dana Faltýnka, Vladimíra Matlacha a dalších.

V naší práci jsme se omezili na určitou délku textu a délku měřených úseků. Analyzovali jsme délky úseků od 2000 tokenů po 5000 tokenů, přičemž ve finální analýze pomocí hierarchického shlukování jsme se zaměřili na délku úseku 5000

tokenů a celkovou délku textu 50000. Výsledky by se pro jiné délky měřených úseků a celkové délky textů mohly významně lišit. Další bádání v této oblasti by se mohlo zaměřit na analýzu toho, jaký vliv má na určení autorství kupříkladu odlišná volba délky úseků, odlišná volba kritérií pro stanovení disperzních tříd, odlišná kritéria pro filtrování, či případně jiný způsob výpočtu nepodobnosti textů. Domníváme se tak, že tato oblast poskytuje mnoho prostoru pro další výzkumy.

5. Reference

Davidson, M., & Sireci, S. (2000). Multidimensional Scalling. *Handbook of Applied Multivariete Statistics and Mathematical Modeling*, 323-352. doi:10.1016/B978-012691360-6/50013-6

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., & Schöch, C. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*.

Faltýnek, D., & Kučera, O. (2022). Parasyntax jako struktura nízko frekventovaných částí textu Hapax legomenon prostředkem textové koheze.

Faltýnek, D., & Matlach, V. (2020). Hapax Remains: authorial features of textual cohesion in authorship attribution. doi:10.13140/RG.2.2.19865.24164

Koppel, M., Schler, J., & Argamon, S. (Leden 2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology, 60*(1), stránky 9-26. doi:10.1002/asi.20961

Mikros, G. K. (2009). Content words in authorship attribution: An evaluation of stylometric features in a literary corpus. *Quantitative Linguistics*, 61-75.

Nielsen, F. (2016). Hierarchical Clustering. V F. Nielsen, *Introduction to HPC with MPI for Data Science* (stránky 195-211). Springer. doi:10.1007/978-3-319-21903-5_8

Plecháč, P. (2021). *Versification and Authorship Attribution.* Prague: Karolinum Press. doi:10.14712/9788024648903

Project Gutenberg. (08. 04 2024). Načteno z Project Gutenberg: https://www.gutenberg.org/

Shade, B., & Altmann, E. G. (2023). Quantifying the Dissimilarity of Texts. *Information, 14*, 271.

Schöch, C. (2017). refcor: Reference corpora for authorship attribution studies. GitHub. Načteno z GitHub: https://github.com/cophi-wue/refcor/tree/master

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary.* Cambridge .

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Addison-Wesley Press.