Filozofická fakulta Univerzity Palackého v Olomouci

Katedra obecné lingvistiky

Analýza využití prefixů při určování autorství textu

bakalářská diplomová práce

Autor: Jan Fiedler

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

Olomouc

2019

Prohlášení

Prohlašuji, že jsem bakalářskou diplomovou práci s názvem Analýza využití prefixů v určování autorství textu vypracoval samostatně a pouze s využitím uvedených zdrojů.

V dne Podpis

Poděkování

Chtěl bych poděkovat Mgr. Danu Faltýnkovi, Ph.D. za jeho vedení, ochotu a pozornost, kterou mi při vypracovávání této bakalářské práce věnoval.

Abstrakt

Název práce: Analýza využití prefixů při určování autorství textu

Autor práce: Jan Fiedler

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

Počet stran a znaků: 40, 72 606

Počet příloh: 19

Abstrakt: Tato bakalářská práce je zaměřena na analýzu využití prefixů při určování autorství textu. V teoretické části je popsáno to, co je forenzní lingvistika, také je zde stručně nastíněna historie tohoto oboru. Dále je zde uvedeno to, co je nejčastěji předmětem zájmu forenzních lingvistů, popsány jsou také dvě základní metody, které forenzní lingvisté využívají – profilování a identifikace autora. Praktická část se zabývá analýzou prefixů v beletristických textech. Je zkoumáno to, zda lze na základě používání předpon odlišit autora, tedy jestli je toto kritérium významné při identifikaci autora. Jsou zde srovnávány texty, které napsal původní autor, s texty autora, který se snažil původního autora napodobit. Cílem práce je tedy zjistit to, zda je mezi oběma autory takový rozdíl v používání jednotlivých prefixů, který by dokázal s vysokou pravděpodobností oba autory odlišit. Vzájemné podobnosti mezi texty jsou graficky zobrazeny a také interpretovány.

Klíčová slova: forenzní lingvistika, určování autorství, prefixy

Abstract

Title: The Analysis of Prefixes Usage in Determining Authorship

Author: Jan Fiedler

Supervisor: Mgr. Dan Faltýnek, Ph.D.

Number of pages and characters: 40, 72 606

Number of appendices: 19

Abstract: This bachelor thesis is focused on the analysis of the use of prefixes in determining the authorship of a text. The theoretical part describes what forensic linguistics is, and briefly outlines the history of this field. Furthermore, the text describes what are the most common interests of forensic linguists, and two basic methods that forensic linguists use – profiling and author identification. The practical part deals with the analysis of prefixes in fictional texts. It is examined whether the use of prefixes can distinguish the author, i.e. whether this criterion is important in author identification. There is a comparison of texts written by the original author with texts by the author who tried to imitate the original author. The aim of this work is to determine if there is such a difference in the use of individual prefixes between the two authors that would be able to distinguish the two authors with a high credibility. The similarities between the texts are graphically displayed and interpreted.

Keywords: forensic linguistics, determination of authorship, prefixes

Obsah

Úvod

V této práci se budu věnovat určování autorství textů. Z tohoto důvodu se v teoretické části zaměřím na stručné vysvětlení toho, co je forenzní lingvistika. V českém pojetí je hlavním cílem forenzní lingvistiky zjistit co nejvíce informací o autorovi pouze na základě jím napsaného textu nebo vyloučit/potvrdit nějakou osobu jako autora anonymního textu.

Existují dvě základní metody, které forenzní lingvisté používají – profilování a identifikace autora (Svobodová 1997). Při profilování je výsledkem analýzy zjištění alespoň některých základních informací o autorovi anonymního textu, např. věk, pohlaví, vzdělání, záliby, povolání apod. Při aplikaci této metody není potřeba žádný jiný text a lingvista veškeré informace získává pouze z anonymního textu. Pro identifikaci autora je ale zapotřebí text srovnávací, o kterém znalec s jistotou ví, že ho napsal autor podezřelý z autorství anonymního textu. Na základě srovnání sporného (anonymního) a srovnávacího textu lingvista určí, zda je autor srovnávacího textu totožný s autorem sporného textu, tedy jestli je podezřelá osoba skutečným autorem anonymního textu.

V praktické části se budu zabývat druhou ze zmíněných metod, tedy identifikací autora. Určování autorství je věnována velká pozornost a stále se hledají nové cesty, jak identifikaci autora zpřesnit a dosáhnout tak vyšší jistoty, že člověk označený jako autor je skutečným autorem.

Při srovnávaní sporného a srovnávacího textu musí být zohledněno velké množství kritérií, např. délka slov, délka vět, frekvence různých slovních druhů atd. Já se v této práci zaměřím na jedno z těchto mnoha možných kritérií. Tímto kritériem bude různé používání prefixů (předpon). Hlavní důvod, proč jsem ke zkoumání autorství vybral právě analýzu prefixů, je ten, že jsem nenalezl žádnou práci nebo studii, která by se zabývala prefixy jako kritériem při určování autorství, a tudíž nevíme, zda může používání slov s různými předponami odlišit autory.

Mým hlavním cílem bude zjistit, zda je rozdíl v používání jednotlivých prefixů u různých autorů takový, že bychom na jeho základě dokázali autory rozlišit, a zda může být toto kritérium při identifikaci autora užitečné. V analýze se zaměřím především na frekvence různých předpon, ale také se budu zabývat tím, jestli identitu autora může prozradit i poměr mezi slovy s předponou a bez předpony.

Z důvodu konzistentnosti při oddělování předpon od zbytku slov se budu řídit pojetím Eleonory Slavíčkové, které je popsáno v Retrográdním morfematickém slovníku češtiny (Slavíčková 1975). K analýze použiji metodu bag-of-words, což je metoda analýzy textu, při které nezáleží na pořadí slov, ale na frekvenci slov (Owsianková et al. 2018). Při analýze budu pracovat s kvantitativně lingvistickým softwarem Quita[1]. Výsledky zobrazené na grafech interpretuji a pokusím se na jejich základě zodpovědět otázku, zda lze na základě používání prefixů odlišit autora.

[1] https://www.quitaonline.com/

I. Teoretická část

1 Co je forenzní lingvistika

Většina definic forenzní (soudní) lingvistiky (někdy i jazyková expertíza) se shoduje na tom, že se jedná o obor, který mimo lingvistiky spadá i do právních věd a kriminalistiky. Například Marie Svobodová definuje forenzní lingvistiku jako „velmi mladý obor aplikované lingvistiky, který se pohybuje na pomezí jazykovědy a právních věd, zejména kriminalistiky a soudnictví." (Svobodová 1997, s. 124).

Existují dvě základní pojetí forenzní lingvistiky. První pojetí je takové, které se zabývá analýzou psaných i mluvených jazykových projevů a výsledky získané touto analýzou dále využívá v kriminalistice. Druhé pojetí zkoumá právnické texty z hlediska jejich srozumitelnosti pro lidi, kteří se nepohybují v právnickém prostředí, a tím pádem pro ně může být obtížné správně porozumět některým právnickým textům, lingvisté se dále podílí na formulaci těchto textů, aby byly pro širokou veřejnost co nejsrozumitelnější (Svobodová 1997, s. 125). V českém prostředí je upřednostňováno spíše první pojetí, tedy takové, které má využití v kriminalistice například při zjišťování autorství anonymních dopisů (Musilová 1995, s. 49).

V českém pojetí je tedy hlavním cílem forenzní lingvistiky určit, kdo je autorem anonymního textu (identifikace autora), nebo alespoň zjistit o autorovi co možná nejvíce základních informací, např. pohlaví, věk, vzdělání atd. (profilování). K identifikaci autora se přistupuje v případě, že máme podezřelou osobu z autorství anonymního textu a je nutné její autorství potvrdit nebo vyvrátit, profilování může pomoct, když není podezření na konkrétní osobu. Těmto dvěma metodám se budeme blíže věnovat v dalších kapitolách.

2 Historie forenzní lingvistiky

Zájem o určení autora textu se začal rozvíjet už v 50. letech 20. století, tento zájem je spojen s písmoznalectvím (také „jazyková expertíza psaných textů"), jehož cílem bylo určení pisatele na základně rukou psaného textu. Mimo expertízy psaných textů se ale objevilo i mnoho dalších oborů, které si za cíl kladly určení autora textu, například jazyková expertíza, technická expertíza písemnosti nebo expertíza písma psacích strojů (Musilová 2005a). Stejný cíl měl i obor zvaný fonoskopie (také forenzní fonetika), ten se ale nesoustředí na psaný text, ale za cíl si klade určení mluvčího na základě mluveného projevu (Musilová 2005a).

Pojem „forenzní lingvistika" byl poprvé použit až v analýze švédského lingvisty Jana Svartvika „The Evans Statements: A Case for Forensic Linguistics" (Svartvik 1968). Zde se Svartvik zabýval případem vraždy z roku 1950, za kterou byl odsouzen k smrti Timothy J. Evans. Evans se přiznal k vraždě své ženy a 14měsíční dcery, ale o tři roky později se k vraždě těchto dvou přiznal i Evansův soused John Christie. Svartvik se proto zabýval analýzou Evansových výpovědí a došel k závěru, že na formulování některých částí výpovědi se musel podílet i někdo jiný než Evans, a tím pádem je Evansova výpověď křivá a Evans byl shledán nevinným a jeho poprava za protiústavní (Svartvik 1968).

2.1 Barák a Neruda

I když se pojem forenzní lingvistka používá až od Svartvikovy analýzy, některé otázky, na které se forenzní lingvisté snaží nalézt odpovědi, byly řešeny už dříve. Často bylo zkoumáno autorství uměleckých textů, je tomu tak například v případě sporného autorství Shakespearova díla nebo díla Alexandra Dumase, ale příklady sporného autorství můžeme nalézt i v českém prostředí, např. spor o Rukopisy nebo pochybnosti nad autorstvím Slezských písní, které je přisuzováno Petru Bezručovi (Musilová 2005a).

Jako příklad sporu o autorství literárního díla můžeme uvést spor o autorství textů (několik básní a jedna povídka), které vyšly v roce 1858 v almanachu Máj podepsány jménem Josefa Baráka. Tyto texty byly Oldřichem Králíkem považovány za dílo Jana Nerudy (Králík 1957).

Barákovo autorství řešil i Pavel Vašák. Ten se zaměřil pouze na Barákovu povídku Kříž pod Petřínem, kterou srovnával s povídkami Jana Nerudy. Vašák měřil délku vět, délku slov, poměr vět se substantivem na posledním místě a vět se substantivem na předposledním místě a poměr slov dlouhých čtyři písmena a slov dlouhých pět písmen (Vašák 1980). Na základě srovnání těchto rysů Nerudu jako autora povídky Kříž pod Petřínem vyloučil. (Vašák 1980).

K této problematice se vrátili i Petr Plecháč a Jiří Flaišman v článku „Problém Barák–Neruda z pohledu současné stylometrie“ (Plecháč a Flaišman 2017), ti odmítli Vašákův závěr a sami došli k závěru, „že sporné básně vykazují velkou míru podobnosti s ranou Nerudovou sbírkou Knihy veršů na rovině lexikální (četnost slovních tvarů, četnost lemmat) i u jednotek obsahujících informace z různých jazykových rovin (frekvence znakových a fonetických n-gramů).“ (Plecháč a Flaišman 2017, s. 766).

3 Forenzní lingvistika v současnosti

Největší rozvoj zažila forenzní lingvistika až v 90. letech, v té době vzniklo několik institucí zabývajících se právě forenzní lingvistikou. V roce 1991 vznikla Mezinárodní asociace forenzní fonetiky a akustiky (The International Association for Forensic Phonetics and Acoustics – IAFPA)[2], o dva roky později pak Mezinárodní asociace forenzní lingvistiky (International Association of Forensic Linguists – IAFL)[3]. V roce 1994 byl založen časopis Forensic Linguistics (od roku 2003 nese název The International Journal of Speech, Language and the Law), který se zabývá problematikou z oblasti forenzní lingvistiky, např. určování autorství, programy pro odhalování plagiátů nebo srozumitelností právních textů.

V mezinárodním kontextu se forenzní lingvistice věnují například Malcom Coulthard, Roger Shuy, Hannes Kniffka nebo John Olsson. V českém prostředí se o rozvoj forenzní lingvistiky zasloužila především Václava Musilová. V České republice se dnes forenzní lingvistikou zabývají hlavně ve Fonetickém ústavu FF UK, kde se soustředí na forenzní fonetiku, dále na univerzitě v Brně a v Plzni (Jurka a Faltýnek 2017).

[2] https://www.iafpa.net/

[3] https://www.iafl.org/

4 Předmět zkoumání forenzní lingvistiky

Forenzní lingvista v praxi přijde do kontaktu převážně s anonymními dopisy. Ale i když jsou anonymní dopisy nejčastějším předmětem zájmu lingvistů, není to ani zdaleka jediný jejich předmět zkoumání. Lingvisté mohou své zkušenosti dále využít při zkoumání autentičnosti deníkových zápisků v případě šikany nebo zneužívání dětí, při určování autora různých právních textů (např. smlouvy, pracovní korespondence atd.), při zkoumání pravosti závěti, při zkoumání odborných textů (autorství článků, pravost znaleckých posudků) nebo při posuzování pravosti svědeckých výpovědí, ale i v mnoha dalších případech (Musilová 2005a, s. 70).

4.1 Anonymní dopisy

Jak už bylo zmíněno výše, nejčastěji forenzní lingvisté zkoumají autorství anonymních dopisů, proto zde stručně popíšu, na jaké typy se anonymní dopisy mohou dělit a jaké informace z nich lze získat.

Druhů anonymních dopisů je velké množství, mohou se lišit obsahem, motivací autora, cílem, rozsahem atd. Další rozdíly v anonymních dopisech jsou zapříčiněny samotnou osobou autora, tedy jeho osobností, příslušností ke konkrétní sociální vrstvě, jeho vzděláním nebo jeho vztahem k adresátovi a dalšími faktory.

Typologie anonymních dopisů je také pestrá, existují výhružné a vyděračské dopisy, dopisy pomlouvačné, urážející, udavačské nebo dopisy podněcující k rasové nebo náboženské nenávisti. Ne všechny anonymní dopisy nutně splňují definici trestného činu, ale i tak může být určení jejich autora užitečné při vyšetřovaní, například kvůli dalšímu svědectví (Musilová 2005a, s. 69). Už z pouhého typu dopisu a jeho obsahu lze leccos zjistit, například dopisy výhružného charakteru, které jsou určeny politikům, jsou nejčastěji napsány muži v důchodovém věku (Musilová 2006, s. 138).

Veškeré závěry plynoucí z analýzy anonymních textů jsou ale založeny pouze na pocitu znalce a je nutné k nim přistupovat jako k neobjektivním. Sami znalci to ve svých závěrech připomínají i formulacemi, které často zní tak, že jde pouze o předpoklad nebo domněnku. Ale i přes tuto nejistotu mohou být závěry při vyšetřování užitečné a mohou dopomoci k nalezení autora textu (Musilová 2006, s. 137). Na správnost předpokladu má také velký vliv dosavadní zkušenost lingvisty se zkoumáním podobných textů.

4.2 Problém předmětu zkoumání

Největším problémem při určování autorství anonymního textu je nedostatek materiálu, dopisy a jiné texty jsou často krátké, a tak neposkytují dostatečné množství informací k určení autora. V krátkých textech může mít autor málo prostoru k realizaci pro něj typických textových vlastností a z toho důvodu nemusí být odlišení jeho textu od ostatních možné. Délka textu je určující pro výběr zkoumaných rysů, pokud má lingvista k dispozici dostatečně dlouhý text, je výhodnější zaměřit se v analýze na kvantitativní rysy (frekvence slov, frekvence různých slovních druhů, délka vět nebo četnost pravopisných chyb), pokud je ale zkoumaný text krátký, je vhodnější se zaměřit na individuální jazykový projev autora (tematická struktura jeho výpovědí) (Jurka a Faltýnek 2017).

5 Metody forenzní lingvistiky

5.1 Profilování

Jak bylo zmíněno dříve, jedním z hlavních cílů forenzní lingvistiky je vytvoření profilu autora, tedy zjištění o něm co možná největší množství základních informací, a to pouze na základě anonymního textu, který má lingvista k dispozici. Proto se k profilování přistupuje ve fázi, kdy mají vyšetřovatelé k dispozici pouze anonymní text a nemají možnost ho srovnávat s texty podezřelé osoby.

Profilování slouží k „zúžení (nebo naopak rozšíření) okruhu osob přicházejících v úvahu jako autoři (pachatelé)“ (Musilová 2005a, s. 68). Díky této metodě, jsme schopni s určitou pravděpodobností určit věk, pohlaví, povolání, vzdělání autora a další obecné informace o něm. Na sestavení autorova profilu se kromě lingvisty často podílí i psycholog, ten pomáhá především při obsahové analýze, tato spolupráce mezi obory vede k přesnějším výsledkům. Spolupráce mezi lingvistikou a psychologii je důvodem, proč je vytváření profilu pachatele na základě textu někdy nazýváno jako psycholingvistické profilování (Musilová 2006).

Informací, které se dají zjistit z pouhého textu je mnoho, kromě obecných údajích o autorovi (věk, vzdělání, pohlaví atd.) například i věrohodnost obsahu nebo reálná nebezpečnost autora. Forenzní lingvisté a psychologové tyto informace mohou získat díky „jazykovým vlastnostem textu, ale i díky samotnému druhu anonymního dopisu, obsahové složce, druhu použité argumentace, věcné správnosti, detailnosti a konkrétnosti sdělení.“ (Musilová 2006, s. 137). To, jaké údaje lze zjistit, závisí na typu anonymního textu, obvykle je totiž možné zjistit pouze některé informace.

Při určování pohlaví autora záleží na druhu anonymního textu, u některých je pro zkušeného forenzního lingvistu určení pohlaví jednoduché, u jiných je i pro znalce s potřebnými zkušenostmi určit pohlaví autora téměř nemožné (Musilová 2006, s. 138). Aby znalec zjistil autorovo pohlaví, zaměřuje se například na obsah textu, zmíněné detaily, na to, jakou argumentaci autor používá a na další vlastnosti textu. Často lze určit pohlaví autora už z pouhého druhu textu, např. výhružné dopisy politikům píší převážně muži.

Věk autora je možné určit díky použitým slovům, tedy na základě slovní zásoby. Dále nám autorův věk může prozradit například výběr z možných variant nebo autorovo vnímání některých prostředků jako příznakových, přičemž autor jiného věku by je mohl vnímat jako bezpříznakové. „Jazykové dovednosti a stylizační návyky z mládí totiž přetrvávají i v dalších životních obdobích a proměňují se méně než jazyk jako celek“ (Musilová 2006, s. 138).

Vzdělání autora lze určit na základě jazykové úrovně autora, ale to nemusí být tak jednoduché, jak by se mohlo zdát na první pohled. Jednoduché je to v případech velmi vysoké nebo naopak velmi nízké jazykové úrovně. Ve většině případů je to ale složitější, protože jazyková úroveň autora nemusí odpovídat jeho vzdělání. „Jsou známy případy, kdy se osoba se základním vzděláním v souvislosti s četbou a kulturními zájmy dokáže v písemné formě vyjadřovat velmi výstižně, zatímco někteří technici dnes mají formulační schopnosti a znalosti pravopisu na úrovni velmi zarážející“ (Musilová 2006, s. 138). Také je potřeba zvážit to, jestli se autor úmyslně nesnaží o snížení svých jazykových dovedností, na takové případy se ale kvůli jeho nedůslednosti obvykle přijde celkem snadno.

Pro zjištění autorova povolání jsou významné především termíny a profesionalismy použité v anonymním textu. K odhalení autorova povolání ale může také pomoct výskyt obratů, které jsou typické pro určité povolání, ale které s obsahem textu nesouvisí (Musilová 2006, s. 139).

To, odkud autor anonymního textu pochází nebo kde žije, lze nejlépe určit z použité slovní zásoby, ne vždy ale mají informace o místní příslušnosti nebo původu autora velký přínos pro vyšetřování (Musilová 2006, s. 139). Například když v dopise, ve kterém autor z něčeho falešně obviňuje svého spoluobčana, budou použity prvky z nářeční oblasti, ve které se autor i adresát nacházejí, nebude informace o místní příslušnosti autora pro jeho identifikaci příliš významná.

Použité lexikum nám může také prozradit příslušnost autora k nějaké sociální skupině nebo vrstvě, nám může prozradit zájmy a záliby autora. V tomto ohledu je významný výskyt slangismů nebo argotismů.

Úkolem lingvistů i psychologů při tvorbě profilu autora je také odhadnout na základě textu jeho důvěryhodnost a reálnou nebezpečnost autora. Zjištění věrohodnosti textu je podstatné například v případě dopisů upozorňujících na trestnou činnost, v dopisech, ve kterých se sám autor k trestnému činu přiznává, nebo v dopisech, jejichž cílem je pomluvit nebo udat někoho. Za věrohodné se dají považovat informace v dopisech, jejichž autor prokazuje znalostí o daném problému, popisuje mnoho detailů a vyjadřuje se spíše věcně. Naopak pokud autor klade důraz na emoce, používá nadsázku a jeho text je plný časových nesrovnalostí, je text označen jako nevěrohodný (Musilová 2006, s. 141).

Při zkoumání toho, jestli může být autor svému okolí opravdovou hrozbou, se lingvisté a psychologové zaměřují nejen na to, co se v dopisu píše, ale i na to, co se tam nepíše, ale v jiných textech stejného druhu je. Většina anonymních dopisů má za cíl pouze zastrašit svou oběť, ale i přesto to mohou mít na oběť dlouhodobý negativní vliv (Musilová 2006, s. 141).

Další informací, kterou je možné získat z anonymního dopisu a která může být důležitá pro nalezení autora, je autorova motivace. Autor ale často svou pravou motivaci neuvádí z důvodu odvedení pozornosti od sebe. Proto je motiv uvedený v dopise často falešný (Musilová 2006, s. 141). Pokud lingvista odhalí pravou autorovu motivaci, může to být významné vodítko při pátrání po autorovi.

5.2 Identifikace autora

Kromě vytvoření profilu autora textu na základě analýzy daného textu je cílem forenzních lingvistů také autorova identifikace. Na rozdíl od profilování, je k identifikaci autora potřebný srovnávací text, který je porovnáván s anonymním textem, a na základě míry shody mezi oběma texty je podezřelá osoba jako autor vyvrácena, nebo potvrzena (pouze s určitou pravděpodobností). Proto se k identifikaci přistupuje v těch případech, kdy je z napsání anonymního dopisu podezřelý konkrétní člověk.

5.2.1 Srovnávací text

K tomu, abychom mohli určit autora textu, je nutné mít i text ke srovnání, kromě sporného (anonymního) textu je tedy za potřebí mít i srovnávací text, který byl napsán konkrétním podezřelým.

Na základě srovnání sporného a srovnávacího materiálu je znalec schopen s určitou pravděpodobností potvrdit nebo vyvrátit autorství podezřelé osoby. Jako srovnávací materiál musí být použit text, který podezřelý napsal, aniž by věděl, že bude později použit pro zkoumání.

Výběr srovnávacích textů je také ovlivněn typem sporného textu. Při určování autorství anonymního dopisu, který je určen nějaké veřejné instituci, je tedy vhodné použít jako srovnávací texty dopisy určené úřadům nebo dopisy týkající se obchodních záležitostí. Naopak když znalec zkoumá soukromý anonymní dopis, je ke srovnávání užitečnější získat soukromou korespondenci podezřelého, jako jsou například jeho osobní dopisy nebo i deníkové zápisky. Kromě výběru správného srovnávacího materiálu je ale také nutné přihlédnout k tomu, v jaké době a situaci podezřelý texty použité ke srovnání napsal.

5.2.2 Předpoklady pro odlišení autorů

Jedním z důvodů, proč vůbec lze na základě textu odlišit autora, je ten, že jazyk poskytuje mnoho příležitostí popsat skutečnost nebo vyjádřit stejné pocity mnoha různými způsoby. Díky této možnosti, kterou nám jazyk poskytuje, je velice pravděpodobné, že se texty dvou autorů liší, a tím pádem dokážeme autory odlišit. Odlišnosti v projevu jsou nejpatrnější v nespisovném jazyce, ale pokud je text dostatečně dlouhý, dostatek možností realizace poskytuje i jazyk spisovný. Pokud je tedy anonymní dopis napsán nespisovným jazykem, minimální délka textu potřebná pro identifikaci autora není tak velká jako v případě dopisů psaných spisovným jazykem. Proto je pro forenzní lingvisty příznivé i to, že většina anonymních dopisů je napsána nespisovným jazykem, a to i mluvčími, kteří spisovný jazyk ovládají (Musilová 2005b, s. 99).

Odlišit autora nám ale kromě možnosti výběru, kterou nám poskytuje jazyk, dopomáhá i rozdílnost jednotlivých autorů. Na vyjadřování každého člověka má vliv mnoho okolností, např. jeho vzdělání, znalosti, osobnost nebo příslušnost ke konkrétní sociální vrstvě. Autor textu má možnost volby použít různé jazykové prostředky „ve všech jazykových rovinách, ale jejich rozsah je různý“ (Musilová 2005b, s. 99).

Na foneticko-fonologické rovině se rozdíly autorů objevují převážně v mluvených projevech, např. regresivní a progresivní asimilace znělosti [sxoda/zhoda], proteze [okno/vokno], úžení [mléko/mlíko] a další hláskové změny. Některé rozdíly se ale mohou vyskytovat i v projevech psaných. Musilová jako příklad odlišností autorů na foneticko-fonologické rovině uvádí výběr z dvojice synonym alespoň/aspoň (Musilová 2005b, s. 100).

Jako příklad odlišnosti mezi jednotlivými uživateli jazyka na morfologické rovině můžeme uvést výběr mezi koncovkami „-ích“ a „-ách“ (o teplácích / o teplákách). Dále třeba koncovku „-é“/„-i“ (Angličané/Angličani).

Na lexikální rovině jsou rozdíly mezi jednotlivými autory nejzřetelnější. Díky velkému množství synonym má totiž autor mnoho možností, jak vyjádřit stejnou skutečnost

různými způsoby. Forenzní lingvista ale musí přikládat jiný význam slovům, která autor použil vědomě a která nevědomě. Mezi ta, která jsou nejčastěji používána nevědomě, patří neplnovýznamové slovní druhy, naopak plnovýznamová slova patří obvykle k těm, která mohou být autorem lehce záměrně pozměněna. „Z praxe vyplynulo, že v této souvislosti jsou velmi důležitým slovním druhem příslovce, jejichž výběr ze synonymních řad je velmi individuální (např. volba mezi zdůrazňujícími slovy zejména/především/hlavně/obzvláště/zvláště/zvlášť)“ (Musilová 2005b, s. 99).

Odlišnosti na syntaktické rovině si autor často sám neuvědomuje, a proto jsou tyto rozdíly pro určení autorství ještě užitečnější než rozdíly na ostatních jazykových rovinách. Rozdíly mezi jednotlivými autory jsou například v délce vět, slovosledu nebo stupněm rozvitosti vět nebo vztahy mezi větami v souvětí (Musilová 2005b, s. 99–100). Rozdíly mezi autory Musilová ilustruje na příkladu souvětí „Jestliže je pěkné počasí, půjdu ven,“ toto souvětí poskytuje autorovi mnoho možností odlišit se od ostatních autorů, např. nahrazení spojky „jestliže“ spojkou „když/pokud…“ (Musilová 2005b, s. 100).

Rozdíly mezi autory se však objevují i v grafické podobě, a to hlavně v pravopisu. odlišnosti mohou být zapříčiněny různým používáním interpunkčních znamének, různým zápisem číslovek nebo množstvím pravopisných chyb. Pravopisná norma ale autorovi neposkytuje mnoho příležitostí vybrat si z více možností realizace, proto velké množství odchylek od normy obvykle znamená nižší vzdělání autora nebo záměrné snižování jazykových schopností.

5.2.3 Jazykový znak

Důležitým pojmem při určování autorství je pojem jazykový znak. Ve forenzní lingvistice se tento pojem používá pro takovou vlastnost textu, která dokáže odlišit jednotlivé autory. Musilová (2005b) uvádí příklad variant češu/češi/česám, kde koncovka -u, -i nebo -ám funguje jako znak, protože má autor na výběr z několika možností, a tím pádem jsme schopni na základě použité koncovky autora odlišit. Jako příklad, co jazykový znak není, Musilová uvádí slovo „říkám“, protože to nedává autorovi možnost výběru koncovky.

Každý jazykový znak je ale typický pro více uživatelů jazyka, proto musíme vzít v potaz mnoho jazykových vlastností textu, abychom odhalili autora s co největší přesností. Takže pokud nalezneme jeden shodný znak ve srovnávacím a sporném textu, neznamená to, že autorem obou byl stejný člověk. Pouze shoda kombinací více jazykových vlastností nám může dopomoct ke správnému určení autora.

Lingvista také musí zvážit míru významu jednotlivých vlastností. Nejvýznamnější jsou takové vlastnosti, které mluvčí používá nevědomě. Z tohoto pohledu jsou ty méně významné znaky v rovině lexikální, naopak k těm významným patří kompozice nebo syntax.

II. Praktická část

6 Postup

Budu analyzovat dva soubory textů – soubor Zrcadla a soubor Trpaslíci. V Zrcadlech je celkem deset textů, osm z nich je napsáno původním autorem, dva jsou podvrhy napsané jiným autorem, tyto podvržené texty se jmenují Kabát a Krkovice. V Trpaslících je osm textů a mezi nimi jsou stejně jakou v Zrcadlech dva podvrhy, jejich názvy jsou Krádež a Takoví jsou hobiti.

Před samotnou analýzou vyberu ze všech textů substantiva, adjektiva a slovesa, následně všechna slova převedu na lemmata, tedy na slovníkový tvar, k tomu využiji online software MorphoDiTa[4], který slouží k morfologické analýze.

Dále vyčlením substantiva, adjektiva a slovesa obsahující předponu. Při určování toho, co předpona je a co není, se budu řídit pojetím Eleonory Slavíčkové, které shrnula v Retrográdním morfematickém slovníku češtiny (Slavíčková 1975).

Následně stručně charakterizuji jednotlivé texty. V charakteristice uvedu nejfrekventovanější slova s předponou i nejfrekventovanější předpony samotné a také poměry mezi slovy s předponou a slovy bez předpony. Pak na základě údajů o poměru mezi slovy s předponou a slovy bez ní odpovím na otázku, zda můžeme odlišit autora pouze podle tohoto kritéria.

Ze slov s předponou zhotovím seznamy podle toho, o jaký slovní druh se jedná. Výsledkem bude celkem šest seznamů, z každého souboru textů tedy vzniknou tři seznamy, které budou obsahovat slova s předponou rozdělená podle slovního druhu. Následujícím krokem bude vytvořit dalších šest seznamů, které budou obsahovat pouze samotné předpony těchto slov, vznikne tedy dalších šest seznamů obsahujících pouze předpony. Celkem tedy vznikne dvanáct seznamů, které podrobím analýze a vzájemnému srovnání.

K analýze všech dvanácti seznamů použiji metodu bag-of-words (BoW)[5], k tomu využiji kvantitativně lingvistický software Quita[6]. Při nastavení parametrů použijí výchozí nastavení Quity pro BoW modely. Výsledky analýz jednotlivých textů se zobrazí na hierarchickém shlukovém grafu, který ukáže, jaké texty si jsou nejblíže.

Poté, co budu mít k dispozici dvanáct grafů, každý založený na jednom z dvanácti seznamů, porovnám je mezi sebou a zhodnotím, jestli jsou od sebe na některém podvrhy a původní texty dostatečně a jednoznačně odděleny. Také zhodnotím, jestli se výsledky liší v závislosti na tom, jaký slovní druh zrovna zkoumáme. Pokud se výsledky budou lišit, vyhodnotím, jaké BoW modely fungují při určování autora nejúspěšněji, jestli ty založené na substantivech, adjektivech, nebo slovesech v případě, že se budou výsledky lišit. Další otázkou bude, jestli k rozlišení autorů lépe poslouží modely založené na celých slovech s předponou, nebo modely založené pouze na předponách.

[4] http://lindat.mff.cuni.cz/services/morphodita/
[5] Metoda analýzy textu, při které nezáleží na pořadí slov, ale pouze na frekvenci slov
[6] https://www.quitaonline.com/

7 Charakteristika jednotlivých textů

V této kapitole stručně charakterizuji jednotlivé texty. V charakteristice se zaměřím především na slova s předponou a předpony samotné, vyjmenuji, jaká slova s předponou a jaké předpony jsou v jednotlivých textech nejfrekventovanější. Ale také uvedu, jakou část textu tvoří zkoumané slovní druhy, tedy substantiva, adjektiva a slovesa, dále se zde zaměřím i na to, kolik slov, která náleží k těmto slovním druhům, má předponu.

Výčet nejfrekventovanějších slov s předponou a samotných předpon nám poslouží později při interpretaci jednotlivých grafů, které budou zobrazovat výsledky BoW modelů. Počet slov s předponou nám dopomůže k zodpovězení otázky, zda lze odlišit autora pouze na základě poměru mezi slovy s předponou a slovy bez předpony.

7.1 Souboru textů Zrcadla

7.1.1 Invektiva proti postmoderním autorům

Nejfrekventovanější substantiva s předponou jsou v tomto textu slova „příběh“, „spiknutí“ a „zdvořilost“, všechna tři se zde vyskytují dvakrát, ostatní substantiva s předponou se neopakují. Nejčastější předpona substantiv je zde předpona „s-“, která je v celém textu pětkrát, dále pak předpony „po-“ a „z-“, přičemž každá z nich je v textu u tří substantiv. Počet všech substantiv v tomto textu je 120, což je asi 25,5 % ze všech slov. Z těchto 120 substantiv jich má předponu 26, což je 21,67 %.

Adjektiva s předponou se zde neopakují vůbec, všechna jsou v textu pouze jednou. Nejfrekventovanější předpony jsou zde u adjektiv předpony „po-“ a „zá-“, obě se v textu vyskytují pětkrát, na dalším místě je pak předpona „pří-“, která je zde u čtyř adjektiv. Celkový počet adjektiv je 39, což je 8,3 % ze všech slov. Adjektiv s předponou je 13, tedy 33,33 % ze všech adjektiv.

Nejfrekventovanější sloveso s předponou je zde slovo „obviňovat“, slovesa „nahradit“ a „prohlašovat“ jsou v textu zastoupena dvakrát, ostatní se v textu neopakují. U sloves se opakuje nejčastěji předpona „roz-“, ta je v textu sedmkrát, předpony „po-“ a „za-“ jsou u pěti sloves. Sloves je v tomto textu 87 (asi 18,51 % ze všech slov), předponu jich má 52,84 %, tedy 46.

7.1.2 Jako rtuť

V tomto textu je nejfrekventovanějším substantivem s předponou slovo „odstín“, vyskytuje se zde čtyřikrát, zároveň je to také jediné substantivum s předponou, které se zde opakuje, všechna ostatní jsou zde zastoupena pouze jednou. U substantiv se nejčastěji opakují předpony „od-“ a „po-“, obě jsou u pěti slov, dále pak předpona „do-“, která je u tří substantiv. Počet substantiv je 211 (30,80 % ze všech slov), z toho 27 jich má předponu, tedy 12,80 %.

Z adjektiv s předponou se opakuje pouze adjektivum „povědomý“, a to dvakrát. Nejfrekventovanější předpony u adjektiv jsou „o-“, „po-“ a „s-“, každá z nich se vyskytuje třikrát. Adjektiv je zde celkem 112 (16,35 % ze všech slov), adjektiv s předponou je zde 19, tedy 16,96 % ze všech adjektiv.

Nejčastěji se opakující slovesa s předponou jsou v tomto textu slova „připomínat“, „pozorovat“ a „protékat“, každé z nich je v textu dvakrát. Z předpon sloves se nejvíce opakuje předpona „po-“, a to čtrnáctkrát, dále pak předpony „u-“, „při-“, „pro-“, „roz-“ a „s-“, všechny tyto předpony jsou zde čtyřikrát. Počet všech sloves je 99, což je 14,45 % ze všech slov. Sloves s předponou je 57, tedy 57,58 %.

7.1.3 Kabát

Tento text je první ze dvou podvržených textů a je napsán jiným autorem než zbylé texty. Zde se ze substantiv s předponovou nejčastěji objevuje slovo „předsíň“, které je zde šestkrát, dvakrát se opakují slova „otázka“, „pohled“, „směr“ a „vzpomínka“, ostatní se neopakují vůbec. Nejfrekventovanější předpona u substantiv je „před-“, která je zde sedmkrát, dále pak předpona „po-“, ta je v textu šestkrát. V textu je 25,53 % substantiv, což je 195 slov, z toho jich má 18,97 % předponu (tj. 37 slov).

Adjektivum s předponou se zde opakuje pouze jedno – „poslední“, to je v tomto textu dvakrát. Předpona, která se u adjektiv vyskytuje nejčastěji, je předpona „po-“, která se opakuje šestkrát. Celkový počet adjektiv v textu je 56 (7,22 % slov z celého textu), předponu jich má 16, tedy 28,57 %.

Nejfrekventovanější sloveso s předponou je slovo „vzpomenout“ s počtem výskytů tři, dále se dvakrát opakují slova „napadnout“, „přijít“, „procházet“, „udělat“, „vstát“, „vypadat“, „zastavovat“, „zavírat“ a „zkoušet“. Z předpon sloves se nejvíce opakuje předpona „po-“, která je v textu jedenáctkrát, a dále předpona „za-“, která je v textu osmkrát. Je zde 151 sloves (19,46 %), z toho jich má předponu 38,41 % (58 slov).

7.1.4 Krkovice

Toto je v souboru textů Zrcadla druhý podvržený text. Zde se ze substantiv s předponou nejčastěji opakuje slovo „povídka“, které se opakuje pětkrát. Nejfrekventovanější předponou substantiv je předpona „po-“, ta se v textu objevuje osmkrát, druhou nejfrekventovanější předponou je předpona „ná-“ zastoupena sedmkrát. Počet všech substantiv je 142 (21,91 % ze všech slov), počet substantiv s předponou je 34, tedy 23,94 %.

Adjektiva s předponou se v textu vyskytují pouze jednou a žádné z nich se neopakuje. Předpona adjektiv, která se objevuje nejčastěji je předpona „po-“, ta je zde šestkrát. Z celého textu jsou 8,33 % slov adjektiva (54 adjektiv), adjektiv s předponou je 18, což je 33,33 % ze všech adjektiv.

Dvěma nejfrekventovanějšími slovesy s předponou jsou „pozřít“ a „zahlédnout“, obě jsou v textu třikrát. Předpona, která se nejčastěji vyskytuje u sloves je předpona „po-“, která je v textu devětkrát, druhá nejčastější předpona je předpona „u-“ spolu s předponou „za-“, obě tyto předpony se opakují sedmkrát. Počet sloves je 140 (21,6 % ze všech slov), z toho 58 jich má předponu, tedy 41,43 %.

7.1.5 Narcismu

V tomto textu je nejfrekventovanější substantivum s předponou je slovo „schod“, druhé nejfrekventovanější je „pohnutí“, ostatní se neopakují. Předpona, která se u substantiv opakuje nejvíce, je předpona „po-“, opakuje se sedmkrát, druhá je předpona „s-“, ta se opakuje šestkrát. Všech substantiv je v textu 136 (27,64 %), z to jich má 30 předponu, což je 22,06 %.

Jediné adjektivum s předponou, které se zde opakuje je slovo „nábožný“, to se tu vyskytuje dvakrát. Předpony „ná-“, „s-“ a „vy-“ jsou u adjektiv dvakrát, ostatní pouze jednou. Počet adjektiv je 40, což je 8,13 % ze všech slov. Adjektiv s předponou je 11, tedy 27,5 %.

Slovesa s předponou se zde opakují pouze dvě – „procházet“ a „uvědomit“ – obě jsou v textu dvakrát. Z předpon sloves je nejfrekventovanější předpona „za-“, která se objevuje sedmkrát, dále pak předpona „vy-“, ta se opakuje pětkrát. Ze všech slov je zde 16,26 % sloves, tedy 80 slov, z toho předponu má 36 slov (45 %).

7.1.6 Nedopalek

V tomto textu se ze substantiv s předponou nejčastěji opakuje slovo „nedopalek“, které je zde celkem sedmnáctkrát, druhé nejfrekventovanější je „vzduch“, to se vyskytuje čtyřikrát. Nejčetnější předpony substantiv jsou předpony „ne-“ a „do-“, které jsou zde devatenáctkrát, dále předpony „po-“ a „vz-“, ty se opakují čtyřikrát. Počet substantiv s předponou je 164, tedy 24,59 % ze všech slov. Předponu z těchto sloves má 63 slov (49,61 %).

Adjektiva s předponou se v tomto textu neopakují, z předpon adjektiv se nejvíce opakuje předpona „o-“, která je zde třikrát. Je zde 51 adjektiv (7,65 % ze všech slov), z toho jich má předponu 11, což je 21,57 %.

Nejfrekventovanějšími slovesy s předponou jsou slovesa „povést“ a „ukázat“, které se v tomto textu objevují třikrát, dvakrát se opakují slovesa „povolit“, „vyrvat“, „začít“ a „zaklesnout“, ostatní se neopakují. Z předpon sloves se nejčastěji opakují předpony „vy-“ a „za-“, a to sedmkrát, dále pak předpony „po-“ a „u-“, ty jsou zastoupeny sedmkrát. Počet všech sloves je 127, tedy 19,04 %. Předponu má 41,43 % sloves (63 slov).

7.1.7 O štěstí

V tomto textu je nejfrekventovanější substantivum s předponou slovo „pokoj“, které se zde opakuje pětkrát, všechna ostatní se objevují pouze jednou. Z předpon substantiv se zde nejčastější vyskytuje předpona „po-“, a to šestkrát. Celkový počet substantiv je 124, což je 26,44 % ze všech slov, předponu jich má 22 (17,74 %).

Z adjektiv s předponou se zde opakuje pouze adjektivum „posedlý“, které je zastoupeno dvakrát. Nejfrekventovanější předpona adjektiv je předpona „po-“, ta se opakuje čtyřikrát. Všech adjektiv je v textu 11,51 %, tedy 54 slov. Adjektiv s předponou je 17, což je 31,48 %.

Sloveso s předponou, které se opakuje nejčastěji je sloveso „najít“, které se opakuje čtyřikrát. Z předpon sloves jsou nejfrekventovanější předpony „na-“, „s-“ a „vy-“, všechny

jsou v textu pětkrát. Počet sloves s předponou je 29, což je 39,19 % ze všech sloves, kterých je 74 (17,78 % ze všech slov).

7.1.8 Oceán

V tomto textu se opakuje pouze jediné substantivum s předponou, a to slovo „náměstí“, které se opakuje třikrát. Nejfrekventovanější předponou substantiv je předpona „ná-“, která se opakuje čtyřikrát, ostatní předpony se neopakují. V textu je celkem 147 substantiv (33,64 % ze všech slov), z toho má 13 předponu (8,84 % slov).

Z adjektiv s předponou se zde neopakuje žádné. Předpony adjektiv, které se opakují, jsou předpony „o-“ a „po-“ opakující se třikrát, dále předpony „roz“, „z-“ a „za-“, které jsou v textu dvakrát. Počet adjektiv je 75, což je 17,16 % ze všech slov. Z těchto 75 adjektiv má 21,33 % předponu (tj. 16 adjektiv).

Opakuje se zde pouze jedno sloveso s předponou, a to sloveso „pozorovat“, které se opakuje dvakrát. Nejfrekventovanější předpony jsou u sloves předpony „po-“, „pro-“ a „za-“, ty se opakují čtyřikrát. Je zde celkem 57 sloves, což je 13,04 % ze všech slov, 25 jich má předponu, tedy 43,86 %.

7.1.9 Společné mlčení

Nejfrekventovanější substantivum s předponou je v tomto textu slovo „náměstí“, které se opakuje čtyřikrát, dále následují slova „námaha“ a „průvodce“ opakující se třikrát. Z předpon substantiv se nejčastěji opakuje předpona „ná-“, a to devětkrát, druhá nejfrekventovanější předpona substantiv je „prů-“, ta se zde objevuje pětkrát. Počet všech substantiv v textu je 242, což je 31,59 % ze všech slov. Substantiv s předponou je 33, tedy 13,64 % substantiv má předponu.

Adjektivum s předponou, které se opakuje nejčastěji je v textu slovo „ohromný“, vyskytuje se zde třikrát. Nejfrekventovanější předponou adjektiv je předpona „po-“, která se zde objevuje šestkrát. Všech adjektiv je v textu 98 (12,79 % ze všech slov), z toho předponu jich má 27, což je 27,55 %.

Ze sloves s předponou jsou nejčastěji se opakující slovesa „doprovázet“, „odrážet“ a „přemýšlet“, všechna tato slovesa jsou zastoupena dvakrát. Předpona, která se u sloves vyskytuje nejvíce, je předpona „po-“ vyskytující se devětkrát, dalšími předponami v řadě jsou předpony „pro-“, „roz-“, „za-“ a „u-“, ty jsou zastoupeny pětkrát. Počet všech sloves je 122, tedy 15,93 % ze všech slov. Z těchto 122 sloves, má 51 sloves předponu, tj. 41,8 %.

7.1.10 Závazek

V tomto textu se ze substantiv s předponou nejvíce opakuje slovo „přísaha“, které se opakuje celkem čtyřikrát, dále pak slovo „závazek“, to se opakuje třikrát. Nejfrekventovanější předpony substantiv jsou „pod-“ a „zá-“, které se zde vyskytují pětkrát, přípona „pří“ pak čtyřikrát. Počet substantiv s předponou je 28, což je 22,05 % ze všech substantiv, kterých je 127, tedy 23,78 % ze všech slov.

Nejfrekventovanějším adjektivum s předponou je slovo „vyhřátý“, to se objevuje dvakrát, zároveň je to jediné adjektivum s předponou, které se v tomto textu opakuje. Z předpon adjektiv se nejvíce opakuje předpona „na-“, která se zde vyskytuje třikrát. Adjektiv je v textu celkem 38 (7,12 % ze všech slov), předponu jich má 9, tedy 23,68 %.

Sloveso s předponou, které se opakuje nejčastěji je sloveso „přemýšlet“, které se v textu vyskytuje pětkrát. Nejfrekventovanější předpony sloves jsou v tomto textu předpony „s-“ a „u-“, vyskytují se zde osmkrát, za nimi následuje předpona „za-“, ta se zde objevuje sedmkrát. Celkový počet sloves je 113, což je 21,16 % ze všech slov. Předponu má 28 sloves, tj. 22,05 %.

7.2 Soubor textů Trpaslíci

7.2.1 Dlouhá cesta

V tomto textu je nejfrekventovanější substantivum s předponou substantivum „nepřítel“, které se opakuje čtyřikrát, druhé nejfrekventovanější je „pozvání“, to se opakuje třikrát. Nejfrekventovanější předpony substantiv jsou dvě: „ne-“ a „po-“, obě se zde objevují šestkrát. Celkový počet substantiv je 177 (tj. 18,93 % ze všech slov), z těch 177 substantiv má 28 substantiv předponu, což je 15,82 %.

Z adjektiv s předponou jsou nejfrekventovanější slova „skutečný“ a „zajímavý“, ta se v textu vyskytují dvakrát, žádná další adjektiva s předponou se neopakují. Předpona, která se zde vyskytuje u adjektiv nejčastěji je předpona „s-“ vyskytující se třikrát, předpony „ne-“, „pří-“, „po“, „u-“, „z-“ a „za“ jsou v textu zastoupeny dvakrát, žádné další přípony se neopakují. Počet adjektiv s předponou je v tomto textu 24, což je 34,78 % ze všech adjektiv (těch je 69, tedy 7,38 % ze všech slov v textu).

Sloveso s předponou, která je v tomto textu nejfrekventovanější, je sloveso „dostat“ společně se slovesem „následovat“, obě se zde vyskytují třikrát. Předpona sloves vyskytující se v textu nejvíce je předpona „u-“, která je u třinácti slov. Další v pořadí jsou předpony „do-“ a „za-“, ty se objevují jedenáctkrát. Sloves je v textu 192, tedy 20,53 %, z toho má 92 předponu, tj. 47,92 %

7.2.2 Krádež

Text Krádež je v souboru Trpaslíci prvním podvrženým textem. Nejfrekventovanější substantivum s předponou je v tomto textu slovo „poklad“, objevuje se zde celkem šestkrát, další v pořadí je až slovo „záležitost“, které je zde pouze dvakrát. Z předpon substantiv je nejfrekventovanější předpona „po-“ vyskytující se desetkrát, druhá nejfrekventovanější je předpona „zá-“, ta je zde třikrát. Celkový počet substantiv je 164, což je 27,29 % ze všech slov. Počet substantiv s předponou je 29, tedy 17,68 % ze všech substantiv.

Adjektiva s předponou, která se jako jediná opakují jsou v tomto textu dvě – „nadcházející“ a „podivný“, obě se opakují dvakrát. Nejfrekventovanější předpona adjektiv je předpona „po“, která je v textu u pěti adjektiv. Adjektiv je v textu 66 (tj. 10,98 % ze všech slov) z toho 17 adjektiv má předponu (25,76 %).

Slovesa s předponou se v tomto textu opakují nejvíce dvakrát, jsou to slova: „dojít“, „dokázat“, „objevit“, „přežít“, „sdílet“ a „zdát“, všechna ostatní slovesa s předponou se objevují pouze jednou. Nejfrekventovanější předponou sloves je zde předpona „po-“ vyskytující se osmkrát, dále předpona „z-“, která je zastoupena sedmkrát, a předpona „do-“, ta se opakuje pětkrát. Celkový počet sloves je 123, což je 20,47 % ze všech slov, z toho má 51 sloves předponu (41,46 %).

7.2.3 Lpění

V tomto textu je nejfrekventovanější substantivum s předponou slovo „schodiště“, to se opakuje třikrát, další v pořadí je slovo „směr“, které se opakuje dvakrát, žádná další se neopakují. Předpona substantiv, která se zde u substantiv vyskytuje nejčastěji je předpona „s-“ s frekvencí pět. Všech substantiv je v textu 104 (22,51 % ze všech slov), z těchto 104 substantiv má 29 substantiv předponu, tj. 17,31 %.

Nejfrekventovanější adjektivum s předponou je zde společně se slovem „proklatý“ slovo „zatracený“, obě se v textu objevují dvakrát. Počet adjektiv je 52 (11,26 % ze všech slov) z toho má 16 adjektiv předponu, což je 30,77 %.

Nejfrekventovanější slovesa s předponou jsou v tomto textu tři, jsou to slovesa „rozhlédnout“, „strávit“ a „vypadat“, všechna se opakují dvakrát. Z předpon sloves je nejfrekventovanější předpona „za-“, která se zde vyskytuje osmkrát, další v pořadí jsou tři předpony vyskytující se čtyřikrát, jsou to předpony „pře-“, „s-“ a „vy-“. Sloves je v textu 83 (19,97 % ze všech slov) z toho jich má 43,37 % předponu (36 sloves).

7.2.4 Nejspíš jsem Vám nerozuměl

V tomto textu je nejfrekventovanější substantivum s předponou slovo „smetí“, vyskytuje se zde čtyřikrát. Nejfrekventovanější předpona substantiv je předpona „s-“, která je v textu u šesti slov, druhá pak předpona „ú“, ta se opakuje čtyřikrát. Substantiv je zde 189 (což je 21,26 % ze všech slov), předponu má 28 substantiv, tedy 14,81 %.

Nejfrekventovanějším adjektivem s předponou je zde slovo „ostatní“, které se opakuje třikrát. Z předpon je u adjektiv nejfrekventovanější předpona „o-“, opakuje se zde pětkrát, druhá je předpona „za-“, ta se u adjektiv opakuje čtyřikrát. Celkový počet adjektiv je zde 87, tj. 9,79 % ze všech slov. Adjektiv, která mají předponu, je 28, což je 28,74 % ze všech adjektiv v textu.

Sloves s předponou, která se zde vyskytují nejčastěji, je osm a opakují se dvakrát, jsou to slovesa „dokázat“, „nabídnout“, „představovat“, „poslouchat“, „prominout“, „vypadat“, „vytahat“ a „zůstat“. Nejfrekventovanější předpona u sloves je předpona „po-“, která se zde opakuje čtrnáctkrát, další pak předpona „vy-“ opakující se jedenáctkrát. Počet všech sloves je 170, tj. 17,12 % ze všech slov. Sloves s předponou je 82, tedy 48,24 % ze všech sloves.

7.2.5 Spáleno na prach

Nejčastěji se vyskytující substantiva s předponou jsou zde dvě – „úsměv“ a „nepřítel“, obě se opakují čtyřikrát. Předpona, která se u substantiv opakuje nejčastěji je předpona „po-“, která se objevuje čtrnáctkrát, další je až předpona „ne-“ s frekvencí šest. Všech substantiv je v textu 212 (22,15 % z celkového počtu slov) z toho jich má 44 předponu, což je 20,75 %.

Nejfrekventovanější adjektivum s předponou je slovo „poslední“, které se opakuje třikrát. Předpona, která se u adjektiv vyskytuje nejčastěji je předpona „po-“, která se opakuje sedmkrát. Celkový počet adjektiv je 97, tedy 10,14 % ze všech slov. Z těchto 97 adjektiv má 35 adjektiv předponu, tj. 36,08 %.

Ze sloves s předponou jsou nejfrekventovanější dvě slovesa, jsou to slova „přijít“ a „pokusit“, která se opakují třikrát. Předpony, které se u sloves opakují nejčastěji, jsou předpony „po-“, „u-“ a „za-“, všechny tyto předpony se u sloves vyskytují desetkrát. Počet sloves v textu je 214, tedy 22,36 % ze všech slov, z toho jich má 91 předponu (42,52 %).

7.2.6 Takoví jsou hobiti

Tento text je napsán jiným autorem než ostatní texty v tomto souboru, a tudíž se jedná po podvrh. V tomto textu se žádné substantivum s předponou neopakuje, nejfrekventovanější předpona u substantiv je předpona „po-“ a „s-“, obě se opakují třikrát. Všech substantiv je zde 130, což je 19,61 % ze všech slov, substantiv s předponou je 11 (8,46 %).

Jediné adjektivum s předponou, které se v tomto textu opakuje je slovo „ztracený“, opakuje se dvakrát. Z předpon adjektiv je zde nejfrekventovanější předpona „z-“, která se opakuje třikrát. Celkový počet adjektiv je 39 (5,88 % ze všech slov) z toho 10 (tj. 25,64 %) má předponu.

Nejfrekventovanější slovesa s předponou jsou slovesa „přijít“ a „zeptat“, obě se opakují třikrát. Předpona vyskytující se u sloves nejčastěji je zde předpona „po-“, která je u osmi sloves. Sloves je v tomto textu celkem 151, což je 22,78 % ze všech slov. Z těchto 151 sloves jich má 58 předponu, tedy 38,41 %.

8 Poměr slov s předponou a bez předpony

V této kapitole shrnu údaje o poměru mezi slovy s předponou a bez předpony podle slovního druhu. Dále se zaměřím na to, jestli jsou nějaké rozdíly mezi slovními druhy a jestli nějaký z nich funguje pro odlišení autora lépe než jiný.

8.1 Zrcadla

V prvním souboru textů Zrcadla jsou z celkem deseti texty podvržené dva – Kabát a Krkovice. Poměr mezi substantivy s předponou a substantivy bez předpony v podvrzích se od původních textů výrazně neliší. V celém souboru má průměrně 18,61 % substantiv předponu, v podvržených textech je to v případě Kabátu 18,97 % a v případě Krkovice 23,94 %. Proto, že se tyto údaje příliš neliší od údajů týkajících se původních textů, nelze podle tohoto kritéria odlišit autora.

Dále se zaměřím na poměr mezi adjektivy s předponou a adjektivy bez ní. Ani u adjektiv se poměr mezi adjektivy s předponou a bez ní v podvrzích výrazně neliší od tohoto poměru v původních textech, a proto nelze ani podle poměru adjektiv s předponou a bez ní odlišit autora.

U sloves je situace zajímavější, v podvrzích je poměr mezi slovesy s předponou a bez ní nižší než u většiny původních textů. V textu Kabát má 38,41 % sloves předponu a v textu Krkovice ji má 41,43 % sloves, ale některé původní texty mají tento poměr podobně nízký, takže ani poměr sloves s předponou a bez ní nelze považovat za kritérium, které by dokázalo jednoznačně odlišit autora.

8.1.1 Trpaslíci

V souboru Trpaslíci jsou stejně jako v Zrcadlech dva podvrhy – Krádež a Takoví jsou hobiti. U substantiv je poměr mezi těmi s předponou a těmi bez ní v podvržených textech různý, v Krádeži má 17,68 % substantiv předponu, což není výrazný rozdíl oproti původním textům, ale ve druhém podvrhu je toto číslo výrazně nižší – pouze 8,46 %. Toto nízké číslo může naznačovat, že autorem tohoto textu je někdo jiný než autor ostatních textů. Ale ani tak to nelze považovat za jednoznačné odlišení autora, protože poměr substantiv s předponou a bez ní v prvním podvrhu se od zbylých textů neliší. Takže v tomto případě je možné na základě poměru mezi substantivy s předponou a bez předpony odlišit autora, ale není tomu tak ve všech případech, jak můžeme vidět v podvrhu Krádež, proto není možně se na toto kritérium v určování autorství spoléhat.

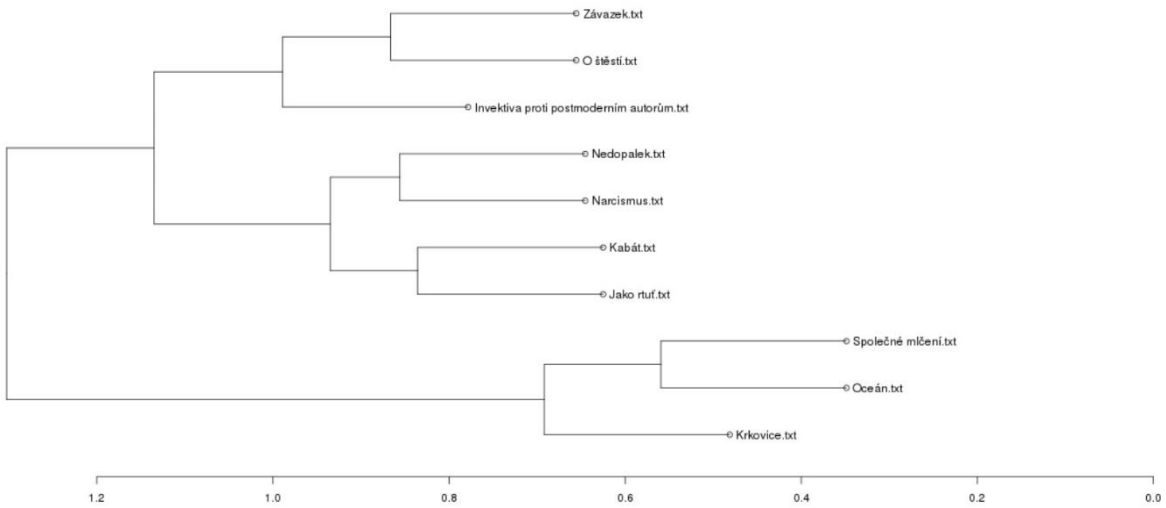
Poměr mezi adjektivy s předponou a adjektivy bez předpony je v souboru Trpaslíci u obou podvržených textů jednoznačně nižší než u původních textů, v podvrhu Takoví jsou hobiti má předponu pouze 25,64 % adjektiv a v podvrhu Krádež je to 25,76 %. Ale protože se nejedná o výrazný rozdíl oproti tomuto číslu v případech ostatních textů, nelze ani toto kritérium prohlásit za takové, jež by dokázalo jednoznačně odlišit autora.

Podobná situace jako u adjektiv je i v případě sloves, i zde je poměr mezi slovesy s předponou a slovesy bez předpony nižší než u původních textů, ale odchylka zde také není tak velká, abychom mohli s jistotou říci, že toto kritérium dokáže odlišit autora.

9 Popis výsledných grafů

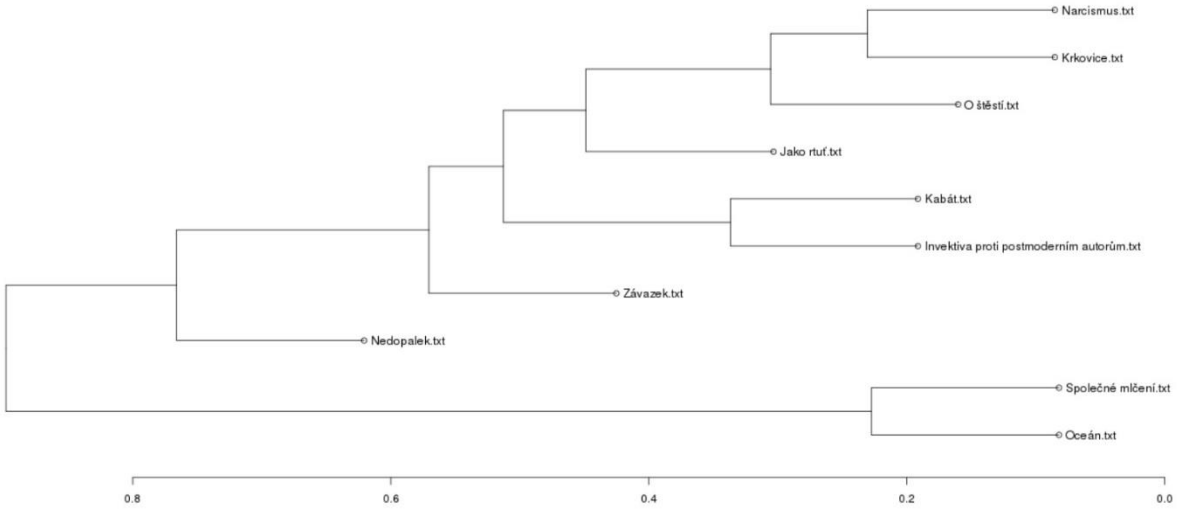
9.1 Zrcadla

V této kapitole se budu věnovat popisu grafů, které zobrazují výsledky BoW modelů získaných díky softwarovému nástroji Quita. Pro analýzu jsem vybral hierarchickou shlukovou analýzu, jejímž výsledkem bude graf, na kterém bude zobrazena vzájemná blízkost textů.



Obrázek 1: Hierarchický shlukový graf, substantiva s předponou (soubor Zrcadla)

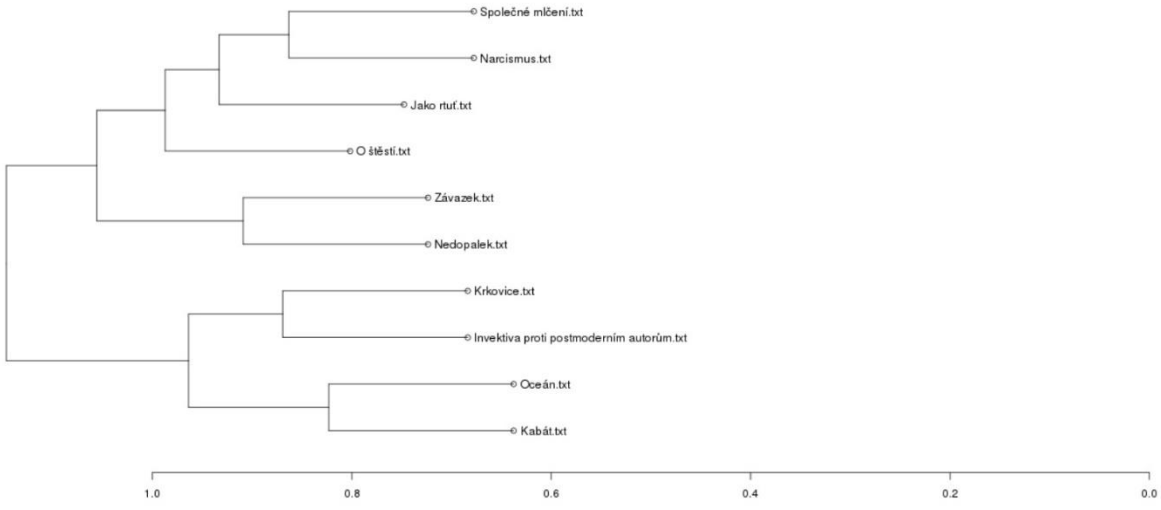
Graf na obrázku 1 ukazuje rozdělení textů na základě substantiv s předponou. Vidíme zde tři texty, které se od ostatních liší nejvíce, jsou to texty Společné mlčení, Oceán a Krkovice. Shoda mezi těmito třemi texty je zapříčiněna vysokou frekvencí slova „náměstí“, v textech Oceán a Společné mlčení je to nejfrekventovanější substantivum s předponou, v textu Krkovice druhé nejfrekventovanější. V ostatních textech nedochází k významné shodě, shoda je zde pouze u slov, které se v textech objevují pouze jednou. Například slovo „nedopalek“ se přesně jednou objevuje v textech Jako rtuť, Kabát, Krkovice, Narcismus a Společné mlčení, proto jsou tři z těchto textů ve stejné skupině. Shoda mezi textem Kabát a Jako rtuť je ve slově „podlaha“ a „zbrojnice“, v obou textech se tato slova vyskytují jednou, dále je spojuje slovo „vzpomínka“, které je v textu Jako rtuť jednou a v textu Kabát dvakrát, v žádném jiném textu se toto slovo neobjevuje.



Obrázek 2: Hierarchický shlukový graf, předpony substantiv (soubor Zrcadla)

Obrázek 2 ukazuje hierarchické rozdělení textů ze Zrcadel na základě předpon substantiv. Ostatním textům jsou nejvzdálenější texty Společné mlčení a Oceán. Tyto dva texty mají stejné nejfrekventovanější substantivum s předponou, je jím slovo „náměstí“, v Oceánu se toto slovo v různých tvarech vyskytuje třikrát, jiná substantiva s předponou se v něm neopakují, ve Společném mlčení se slovo „náměstí vyskytuje čtyřikrát“. Oba tyto texty mají i shodnou nejfrekventovanější předponu, a to právě předponu „ná-“ ze slova „náměstí“. V Oceánu se tato předpona objevuje čtyřikrát a ve Společném mlčení dokonce devětkrát, to je zapříčiněno nejenom častým opakováním slova „náměstí“, ale také opakováním druhého nejfrekventovanějšího substantiva s předponou, kterým je slovo „námaha“.

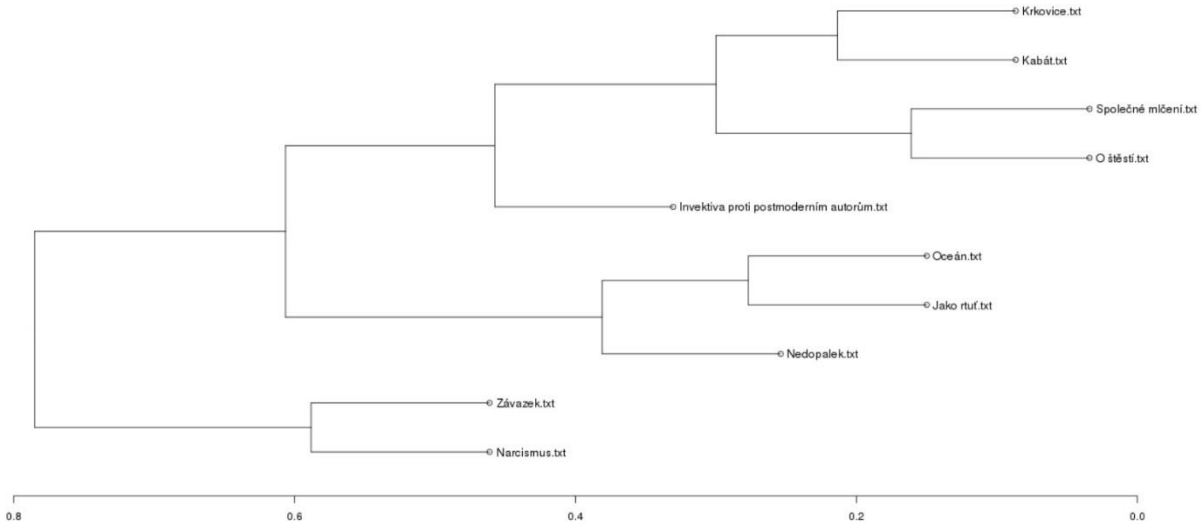
Texty, které byly napsány jiným autorem (Kabát a Krkovice), analýza na základě frekvence jednotlivých předpon substantiv neodděluje od ostatních. Nejfrekventovanější předponou Krkovice je předpona „po-“, ta se v textu vyskytuje u osmi substantiv, značná část těchto výskytů je dána slovem „povídka“, které je zde v různých tvarech zopakováno pětkrát. Předpona „po-“ je nejfrekventovanější předponou substantiv i v textech Narcismus a O štěstí, nejfrekventovanější je i v textu Jako rtuť, zde se ale o první místo dělí s předponou „od-“. V druhém podvrženém textu je nejfrekventovanější předponou předpona „před“, která se v něm objevuje u sedmi substantiv, tato předpona není ve větším počtu zastoupena v žádném z textů. To, co spojuje tento text s textem Invektiva proti postmoderním autorům je předpona „s-“, v textu Kabát je tato předpona s pěti výskyty až třetí nejfrekventovanější předponou po předponách „před-“ a „po-“, ale v textu Invektiva proti postmoderním autorům je to se shodným výskytem jako v Kabátu nejfrekventovanější předpona substantiv. Specifický je v tomto případě text Nedopalek, zde jsou nejfrekventovanější předpony „ne-“ a „do-“, ty se zde vyskytují devatenáctkrát, tento vysoký počet výskytu je zapříčiněn častým opakováním slova „nedopalek“, které se v textu vyskytuje také devatenáctkrát.



Obrázek 3: Hierarchický shlukový graf, adjektiva s předponou (soubor Zrcadla)

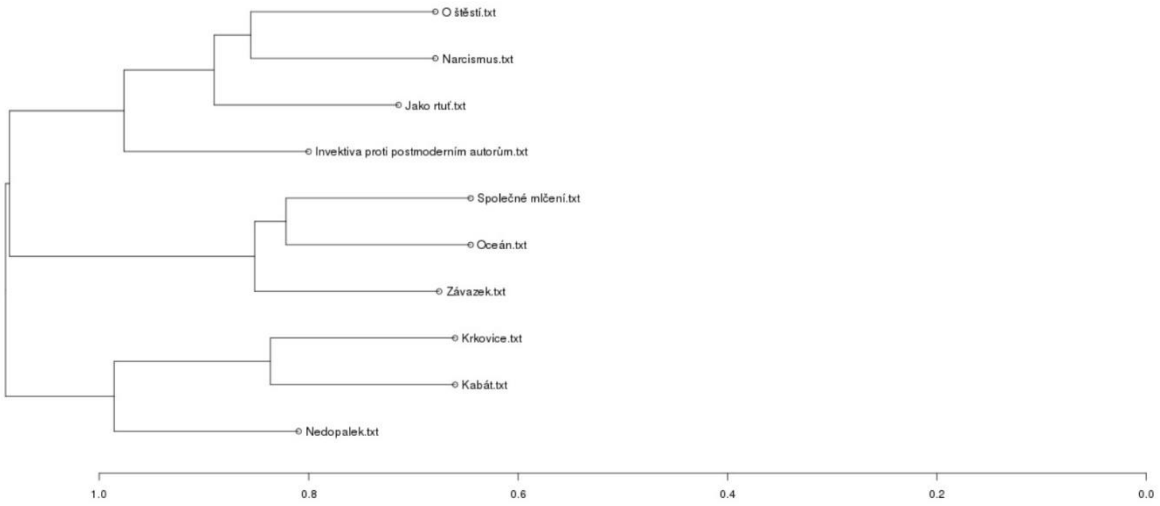
Na grafu z obrázku 3 vidíme rozdělení textů podle adjektiv s předponou. Podvržené texty jsou odděleny od většiny ostatních, ale jsou ve skupině s texty Invektiva proti postmoderním autorům a Oceán. V těchto čtyřech textech se žádná adjektiva s předponou neopakují

vícekrát než jednou, kromě textu Kabát, ve kterém se dvakrát opakuje slovo „poslední“, to je ale jediné adjektivum, které se zde opakuje. Shoda mezi textem Krkovice a Invektiva proti postmoderním autorům je dána společným výskytem slova „pohozený“ a „složitý“, kromě těchto dvou textů se tato slova v žádném jiném textu neobjevují.



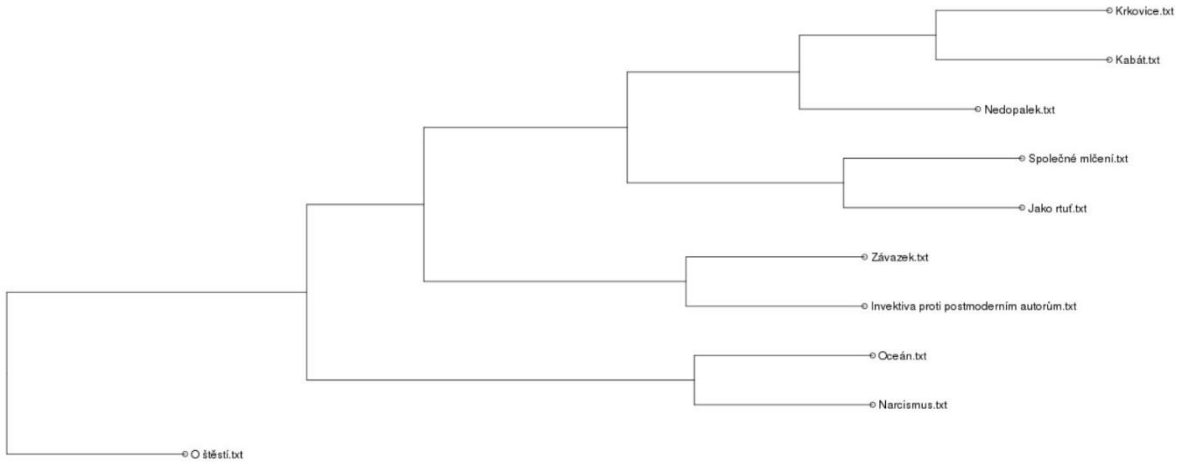
Obrázek 4: Hierarchický shlukový graf, předpony adjektiv (soubor Zrcadla)

Na obrázku 4 je znázorněno rozdělení textů na základě výskytu předpon adjektiv. Můžeme zde vidět, že nejvíce se od ostatních textů liší texty Závazek a Narcismus, tato shoda je zapříčiněna výskytem předpony „po-“, ta se v obou textech objevuje pouze jednou, ve všech ostatní textech je nejméně dvakrát, většinou však ještě častěji. Dále se v těchto dvou textech u adjektiv vyskytuje dvakrát předpona „vy-“, v Závazku je předpona „vy-“ dvakrát kvůli adjektivu „vyhřátý“, které je zde také dvakrát, v Narcismu tato předpona pochází z jiných slov. Na tomto grafu také vidíme, že se oba podvržené texty sdružují ve stejném shluku, přesto ale mají blízko k původním textům. Podvržené texty mají společnou nejfrekventovanější předponu, tou je předpona „po-“, která se v obou vyskytuje šestkrát. Předpona „po-“ je ale nejfrekventovanější i v textech O štěstí a Společné mlčení, popřípadě se o pozici nejfrekventovanější předponu dělí s jinou předponou jako v textech Invektiva proti postmoderním autorům a Oceán.



Obrázek 5: Hierarchický shlukový graf, slovesa s předponou (soubor Zrcadla)

Graf na obrázku 5 ukazuje vzájemnou blízkost textů na základě sloves s předponou. Oba podvržené texty jsou spolu s textem Nedopalek odděleny od ostatních textů, ve všech těchto třech textech se objevuje sloveso „udělat“, to se kromě těchto textů vyskytuje už jen ve Společném mlčení. Původní text Nedopalek má s textem Kabát společný výskyt slovesa „ohnout“, které není zastoupeno v žádném jiném textu. S Krkovicí je Nedopalek spojen kvůli slovu „vycházet“, které se nikde jinde neobjevuje, a slovu „začít“, které se vyskytuje už jen v textu Společné mlčení. V textech Krkovice a Kabát se shodně vyskytují slovesa „otevřít“, „vznášet“, „zahlédnout“ a „napsat“, ta nejsou v žádných dalších textech. Tyto podvržené texty mají i společný výskyt slov „napadnout“ a „přijít“, tato dvě slova se ale objevují i v jednom dalším textu – „napadnout“ v Jako rtuť a „přijít“ v Oceánu.



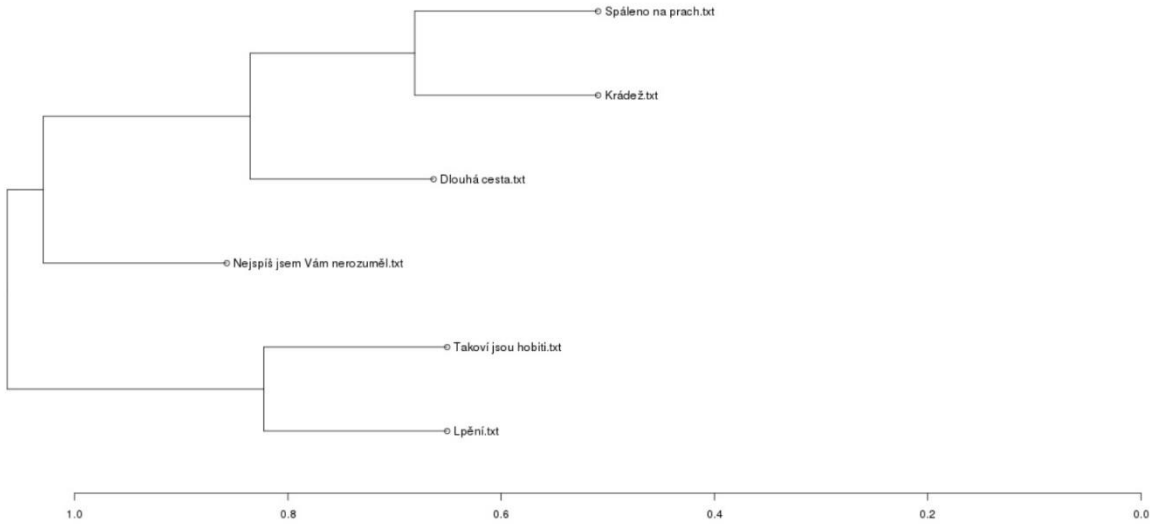
Obrázek 6: Hierarchický shlukový graf, předpony sloves (soubor Zrcadla)

Na grafu z obrázku 6 vidíme, jak jsou texty rozděleny na základě předpon sloves. Výrazně se od všech ostatních textů liší text O štěstí, to je zapříčiněno třemi nejfrekventovanějšími předponami sloves v tomto textu, což jsou předpony „na-“, „s-“ a „vy-“. S výjimkou textu

Závazek, ve kterém je předpona „s-“ nejfrekventovanější, nejsou tyto předpony na prvních místech, co se frekvence týče. V ostatních textech jsou tedy předpony „na-“, „s-“ a „vy-“ spíše méně zastoupené. Předpona „na-“ je zde často kvůli častému opakování slovesa „najít“, výskyt předpon „s-“ a „vy-“ je dán různými slovesy, např. „skládat“, „složit“, „stáčet“, „stěžovat“ nebo „vyhnat“, „vyškrábat“, „vyplňovat“ atd.

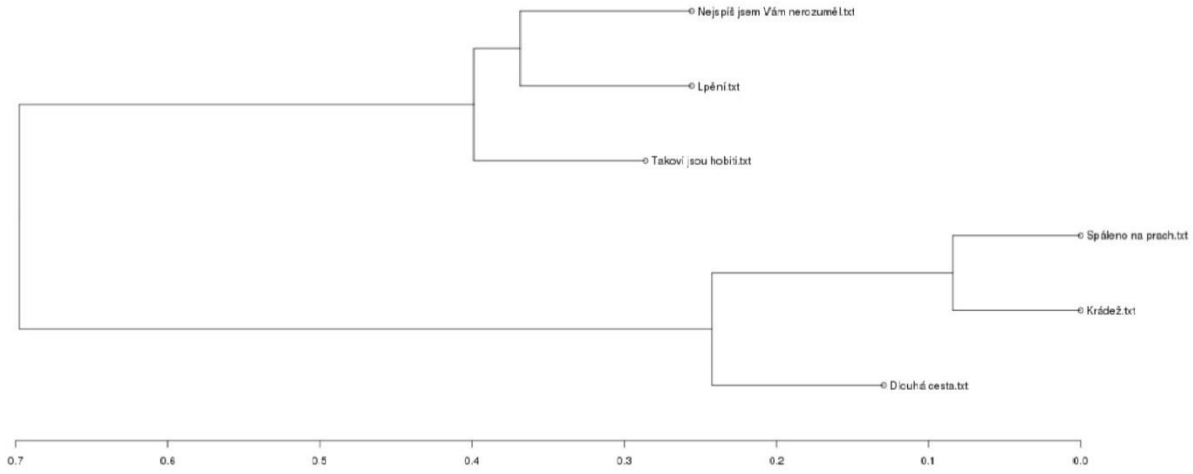
Stejně jako na obrázku 5, který ukazuje příbuznost textů na základě sloves s předponou, zde vidíme, že se oba podvrhy sdružují u sebe. Toto sdružení je zapříčiněno stejnou nejfrekventovanější předponou, kterou je předpona „po-“, ale také druhou nejfrekventovanější předponou, kterou je předpona „za-“. Předpona „po-“ je často nejfrekventovanější předpona, je tomu tak i v textech Jako rtuť, Oceán a Společné mlčení. Text Společné mlčení má navíc shodnou i druhou nejfrekventovanější předponu s podvrhy Kabát a Krkovice. V textu Nedopalek se také předpony „po-“ a „za-“ pohybují podle počtu výskytu na dvou nejvyšších příčkách, ale v opačném pořadí než v podvrzích. Předpona, která také spojuje oba padělky, je předpona „ote-“, ta v obou textech pochází se slovesa „otevřít“ a v žádném jiném textu se nevyskytuje. Dále je pak spojuje výskyt předpony „v-“, která se v obou podvrzích objevuje dvakrát, v ostatních textech se objevuje pouze jednou nebo vůbec.

9.2 Trpaslíci



Obrázek 7: Hierarchický shlukový graf, substantiva s předponou (soubor Trpaslíci)

Obrázek 7 zobrazuje hierarchické rozdělení textů na základě substantiv s předponou. Můžeme zde vidět, že jeden z padělků – Takoví jsou hobiti – se společně s původním textem Lpění oddělil od ostatních, u druhého padělku (Krádež) není odlišení patrné. Texty Lpění a Takoví jsou hobiti spojuje výskyt slov „pohled“ a „směr“. Druhý podvrh, tedy text Krádež, je s textem Spáleno na prach sdružený kvůli výskytu slova „poklad“, toto slovo se neobjevuje v žádném z dalších textů. Další slova, která tyto texty spojují jsou substantiva „nepřítel“, „úsměv“, „pomoc“ a „pozornost“, ta se ale na rozdíl od slova „poklad“ objevují i v jiných textech.

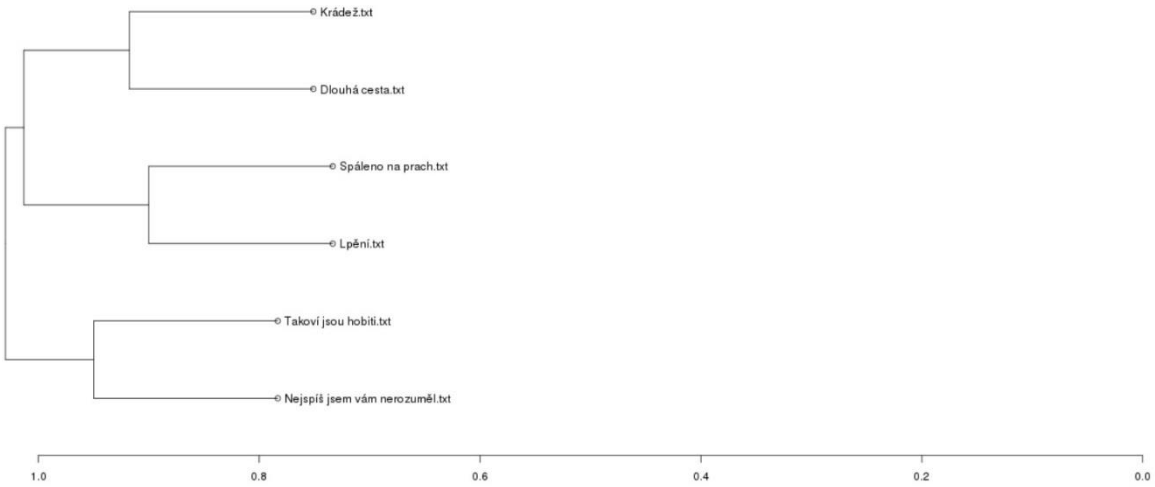


Obrázek 8: Hierarchický shlukový graf, předpony substantiv (soubor Trpaslíci)

Na obrázku 8 vidíme, že se texty na základě předpon substantiv rozdělily do dvou skupin, každý podvrh je v jedné z nich.

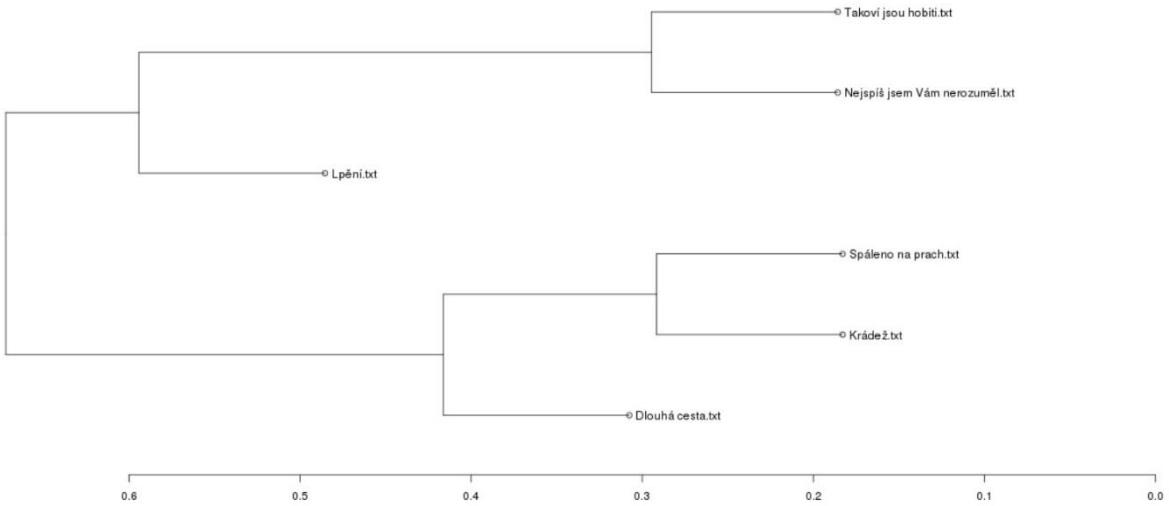
To, co spojuje první skupinu textů, ve které jsou texty Nejspíš jsem Vám nerozuměl, Lpění a Takoví jsou hobiti, je vysoká frekvence předpony „s-“, ve všech těchto textech je to nejfrekventovanější předpona substantiv, ve zbývajících textech je frekvence této předpony výrazně nižší. V textu Nejspíš jsem Vám nerozuměl je tento vysoký výskyt předpony „s-“ zapříčiněn opakováním slov „smetí“, v textu Lpění pak díky slovům „schodiště“ a „směr“ a v podvrhu Takoví jsou hobiti se předpona „s-“ objevuje vždy u jiného slova.

Texty ve druhé skupině spojuje předpona „po-, ta je ve všech těchto třech textech nejfrekventovanější. V Dlouhé cestě se na výskytu této předpony z velké části podílí slovo „pozvání“, které je zde druhé nejfrekventovanější, v Krádeži přes polovinu výskytů předpony „po-“ zapříčiňuje nejfrekventovanější substantivum s předponou „poklad“ a v textu Spáleno na prach je vysoký počet této předpony určen např. slovy „pohled“, „poklad“, „postel“ a „poznání“, která se všechna opakují v textu dvakrát.



Obrázek 9: Hierarchický shlukový graf, adjektiva s předponou (soubor Trpaslíci)

Graf na obrázku 9 znázorňuje rozdělení textů podle adjektiv s předponou, vidíme zde, že je od ostatních textů oddělen jeden z podvržených textů spolu s textem Nejspíš jsem Vám nerozuměl. To je zapříčiněné adjektivem „bezpečný“, které se objevuje v obou těchto textech, ale v žádném dalším. Druhý plagiát je ve skupině s ostatními texty, nejblíže má k Dlouhé cestě, to je dáno adjektivem „zajímavý“, které se pouze v těchto dvou textech.



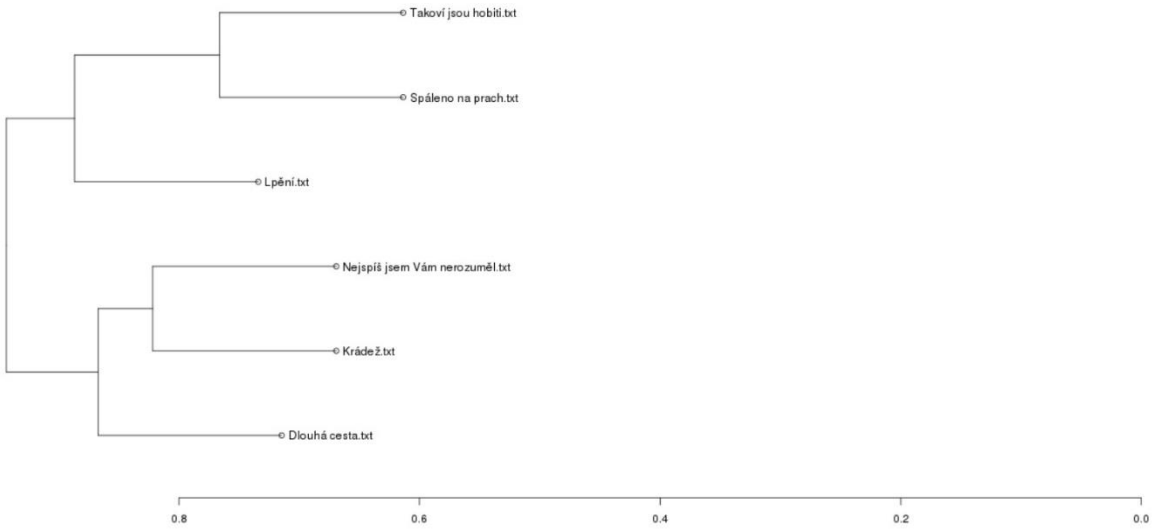
Obrázek 10: Hierarchický shlukový graf, předpony adjektiv (soubor Trpaslíci)

Na obrázku 10 vidíme rozdělení textů podle předpon adjektiv. Výrazně oddělen není ani jeden z podvrhů.

Podvrh Takoví jsou hobiti má blízko k textu Nejspíš jsem Vám nerozuměl. Tato shoda je zapříčiněna společným výskytem předpony a „v-“, ta se objevuje pouze v těchto dvou textech. Další předpony, které se vyskytují v obou textech jsou předpony „bez-“, „na-“, „dů-“ nebo „z-“, ty se ale vyskytují i v jiných textech.

Shoda druhého podvrhu, Krádeže s textem Spáleno na prach je dána stejnou nejfrekventovanější předponou, tou je předpona „po-“. Tato předpona je v textu Spáleno na prach

nejfrekventovanější kvůli výskytu slova „poslední“, které se zde opakuje třikrát, a slova „poházený“, to se v tomto textu opakuje dvakrát. V textu Krádež se z adjektiv s touto předponou opakuje pouze slovo „poslední“, a to dvakrát. V těchto textech je předpona „po-“ u adjektiv nejfrekventovanější předpona, zatímco v jiných textech se vyskytuje minimálně. Další shoda mezi těmito texty je zapříčiněna předponou „nad-“, která je přítomna pouze v těchto textech.

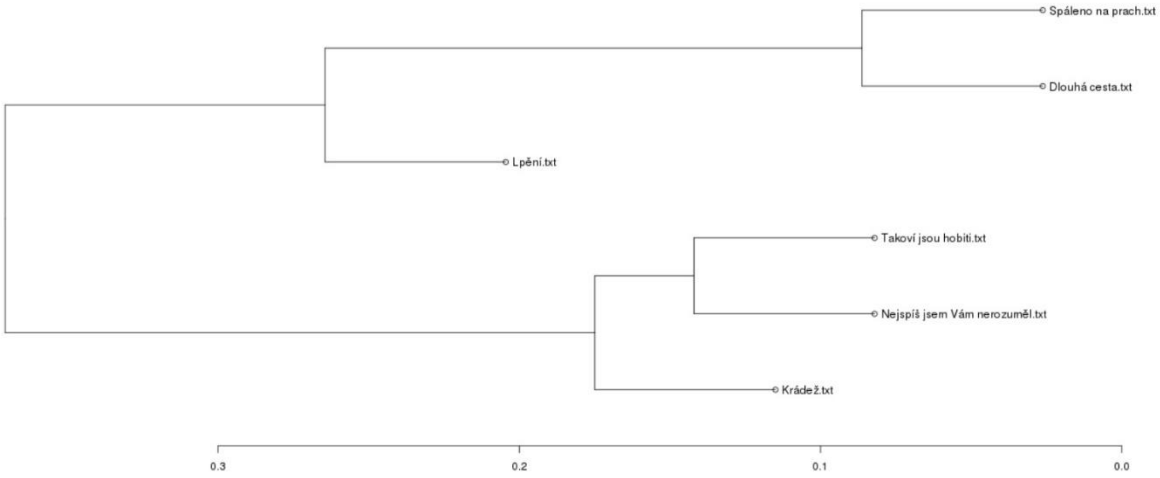


Obrázek 11: Hierarchický shlukový graf, slovesa s předponou (soubor Trpaslíci)

Graf na obrázku 11 ukazuje rozdělení textů na základě sloves s předponou. Můžeme na něm vidět, že všech šest textů je rozděleno do skupin po třech, přičemž v každé z nich je jeden podvrh.

Texty Takoví jsou hobiti a Spáleno na prach mají společné nejfrekventovanější sloveso s předponou, kterým je sloveso „přijít“, v obou textech se vyskytuje třikrát. Společné jsou také slova „naplnit“, „potřebovat“, všechna tato tři slovesa se vyskytují pouze v těchto dvou textech, v žádném jiném.

Text Krádež má blízko k textu Nejspíš jsem Vám nerozuměl díky slovesům „překvapit“ a „ztuhnout“, tato dvě slovesa se objevují pouze v těchto dvou textech. Další slovesa, která jsou v obou textech jsou „dojít“, „dokázat“, „promluvit“, „vypadat“, „zdát“ a „ztratit“, tato slovesa se ale vyskytují i v jiných textech.



Obrázek 12: Hierarchický shlukový graf, předpony sloves (soubor Trpaslíci)

Obrázek 12 zobrazuje, jak jsou od sebe texty odděleny na základě předpon sloves. Oba podvrhy se od většiny původních textů oddělily, je mezi nimi ale text Nejspíš jsem Vám nerozuměl.

Všechny tyto tři texty mají společné to, že nejfrekventovanější předpona je „po-“, ta je nejfrekventovanější i v textu Společné mlčení, tam se ale o první místo dělí s předponami „u-“ a „za-“. V textech Takoví jsou hobiti, Nejspíš jsem Vám nerozuměl a Krádež je předpona „po-“ jediná nejfrekventovanější předpona. Její výskyt je v textu Takoví jsou hobiti dán slovem „potřebovat“, to se zde vyskytuje dvakrát, ostatní slovesa s předponou „po-“ jsou zde pouze jednou, jsou to například slova „pokoušet“, „potřást“ nebo „poznat“. V textu Nejspíš jsem Vám nerozuměl je předpona „po-“ nejfrekventovanější díky užitím slova „poslouchat“, které se v textu vyskytuje dvakrát, a dále slovy, které se zde objevují pouze jednou, např. slova „počkat“, „popadat“, „poradit“ a „povolit“. V Krádež je výskyt této předpony ovlivněn slovy „pochopit“, „podívat“, „pohnout“, „popírat“, „počítat“, „považovat“ atd., přičemž všechna tato slova se zde vyskytují pouze jednou.

To, co spojuje texty Takoví jsou hobiti a Nejspíš jsem Vám nerozuměl je předpona „vy-“, ta se v Krádeži objevuje u sloves pouze jednou, v Nejspíš jsem Vám nerozuměl je to ale druhá nejfrekventovanější předpona a v textu Takoví jsou hobiti třetí nejfrekventovanější. V nejspíš jsem Vám nerozuměl se předpona „vy-“ objevuje pouze u slov, která se neopakují, např. slovesa „vybrat“, „vyhnout“ nebo „vyklepat“. Třetí nejfrekventovanější předpona v Takoví jsou hobiti je to díky slovu „vypadat“, které se opakuje dvakrát a dále díky slovům, která se neopakují vůbec jako slovesa „vybrat“ nebo „vydechnout“

Další předpona, která spojuje texty Nejspíš jsem Vám nerozuměl a Takoví jsou hobiti je předpona „pro-“, která je v textu Nejspíš jsem Vám nerozuměl čtvrtá nejfrekventovanější a v textu Takoví jsou hobiti třetí nejfrekventovanější. Tato předpona se v obou textech vyskytuje pouze u sloves, která se neopakují. Jediná výjimka je slovo „prominout“, to se opakuje dvakrát v textu Nejspíš jsem Vám nerozuměl.

10 Srovnání výsledných grafů

10.1 Zrcadla

10.1.1 Grafy založené na celých slovech s předponou

V grafu na obrázku 1, který znázorňuje příbuznost textů podle celých substantiv s předponou, se od ostatních textů oddělila skupina textů Společné mlčení, Oceán a Krkovice, z této skupiny je podvrh pouze jeden text (Krkovice), zbylé dva (Společné mlčení a Oceán) jsou napsány původním autorem. V tomto případě tedy nelze na základě substantiv s předponou s jistotou odlišit autora.

Rozdělení podle celých adjektiv s předponou je znázorněno na obrázku 3, zde se od zbytku textů oddělily oba podvrhy, ale jsou ve skupině ještě s dalšími dvěma texty (Invektiva proti postmoderním autorům a Oceán). Oba podvrhy jsou zde ve stejné skupině, ale jsou i ve stejné skupině s většinou původních textů, podle adjektiv s předponou tedy také nelze s jistotou určit, kdo je autorem daného textu, stejně jako v případě substantiv s předponou.

Na grafu z obrázku 5, který ukazuje příbuznost textů na základě celých sloves s předponou, můžeme vidět, že se oba podvrhy (společně s jedním původním textem) oddělily od všech původních textů. Výsledky založené na výskytu sloves s předponou jsou tedy přesnější než výsledky založené na výskytu substantiv a adjektiv s předponou, přesto nejsou výsledky zcela jednoznačné, protože spolu s podvrhy byl od zbytku původních textů i jeden text, který byl napsán původním autorem.

10.1.2 Grafy založené na předponách

Obrázek 2 zobrazuje příbuznost textů podle výskytu samotných předpon u substantiv. Podle tohoto kritéria se od ostatních textů nejvíce liší texty Společné mlčení a Oceán. Oby tyto texty jsou napsány původním autorem, toho kritérium tedy nelze použít k odlišení autora.

Na grafu z obrázku 4 vidíme, že rozdělení textů podle předpon adjektiv také nedokáže odlišit autory textů. Dva původní texty jsou si navzájem sice nejblíže, ale oba jsou ve skupině s ostatními původními texty. Dva texty, které se podle předpon adjektiv od zbytku liší nejvíce, jsou Závazek a Narcismus, tedy texty od původního autora.

Rozdělení textů podle předpon, které se vyskytují u sloves (obrázek 6), je z hlediska odlišení autora podobné, jako rozdělení podle předpon adjektiv. Oba podvrhy jsou si blízké, ale zároveň jsou ve skupině s dalšími texty, které jsou napsány původním autorem, takže také nelze jednoznačně určit autora.

10.1.3 Shrnutí

V souboru textů Zrcadla nedokázal žádný BoW model založen na předponách odlišit texty původního autora od jiného autora. BoW modely předpon adjektiv a sloves sice ukázaly vzájemnou blízkost obou podvrhů, ale dostatečně je neoddělily od původních

textů. BoW model založen na předponách substantiv pak neukázal ani vzájemnou shodu mezi oběma podvrhy.

BoW modely založené na celých slovech s předponami dokázaly odlišit původní texty od podvržených úspěšněji než modely založené pouze na předponách, ale ani tak nejsou výsledky jednoznačné. Nejlépe zde fungoval BoW model založený na slovesech, ten dokázal shodu mezi oběma podvrhy, ale ani ten je nedokázal věrohodně odlišit od původních textů, protože do stejné skupiny přiřadil i původní text Nedopalek.

Shoda mezi podvrhy je dána hlavně společným výskytem slov „otevřít“, „vznášet“, „zahlédnout“ a „napsat“, což je pravděpodobně pouze důsledek popisovaných skutečností v jednotlivých povídkách a použití slov s jinými předponami by významně změnilo smysl textu, tudíž autor neměl dostatečnou volnost při volbě slov. Právě proto, že je autor ve volbě slov omezen tím, co chce říct a o čem daná povídka je, je toto kritérium pro určení autora nepodstatné a shoda je tedy víceméně náhodná.

10.2 Trpaslíci

10.2.1 Grafy založené na celých slovech s předponou

Rozdělení textů podle výskytu celých substantiv s předponou, které můžeme vidět na grafu z obrázku 7, ukazuje, že se nejvíce od zbylých textů dvojice Takoví jsou hobiti a Lpění, z toho je pouze jeden text podvrh, druhý podvrh je ve skupině s původními texty. Podle substantiv s předponou tedy v tomto případě nelze odlišit autora.

Ani BoW model založený na adjektivech s předponou nedokázal přesvědčivě oddělit původní texty od podvrhů. Stejně jako v předchozím případě je jeden z podvrhů oddělen od většiny původních textů, ale s jedním z nich je ve stejné skupině, druhý podvrh je ve skupině mezi většinou původních textů.

Stejně jako modely založené na celých substantivech a adjektivech nedokázal spolehlivě oddělit původní texty od podvrhů ani model založený na slovesech s předponou (graf na obrázku 11). Zde se všech šest textů rozdělilo do dvou skupin po třech, přičemž v každé je jeden podvrh a dva texty napsané původním autorem.

10.2.2 Grafy založené na předponách

Na obrázku 8, který ukazuje příbuznost textů na základě předpon substantiv, vidíme, že ani tentokrát nebyly podvrhy odděleny od původních textů. Každý z podvrhů byl přiřazen do skupiny mezi dva texty napsané původním autorem.

BoW model založený na předponách adjektiv (obrázek 10) je téměř shodný s BoW modelem, který je založený na předponách substantiv. I zde jsou texty rozděleny do dvou skupin, přičemž ke každé z nich je jeden podvrh a dva původní text, takže ani na tomto grafu nejsou odděleny původní texty od podvrhů.

Z modelů založených na předponách dokázal nejpřesněji oddělit původní a podvržené texty model založený na předponách sloves (obrázek 12). Tento model rozdělil texty do dvou skupin po třech textech, v jedné skupině jsou tři původní texty a ve druhé jsou oba

dva podvrhy s jedním původním textem. Právě proto, že je jeden původní text ve skupině s oběma podvrhy, nelze ani toto kritérium prohlásit za spolehlivé při určování autorství.

10.2.3 Shrnutí

V souboru textů Trpaslíci byly BoW modely založené na předponách přesnější než modely založené na celých slovech, což je rozdíl oproti souboru textů Zrcadla, kde byly úspěšnější modely založené na celých slovech. Nejpřesněji zde fungoval model, který byl založen na předponách sloves, přesto ani ten nedokázal texty oddělit jednoznačně. Nicméně oba soubory textů mají společné to, že nejpřesněji dokázaly původní a podvržené texty oddělit modely založené na slovesech, i když v souboru Zrcadla to byla celá slovesa a v souboru Trpaslíci předpony sloves.

V tomto BoW modelu založeném na předponách sloves, který na souboru Trpaslíci fungoval nejpřesněji, bylo oddělení podvrhů od většiny původních textů dáno výskytem předpony „po-“. Předpona „po-“ byla v obou podvrzích a jednom původním textu ta nejfrekventovanější. Slova s touto předponou nemají v těchto třech textech frekvenci vyšší než dvě, nepatří tedy k těm nejfrekventovanějším, a proto se výsledek liší od modelu, který byl založen na celých slovech. I v tomto případě se zdá, že shoda mezi texty byla spíše náhodná stejně jako u souboru Zrcadla, kde byl úspěšnější model založený na celých slovech.

Závěr

Cílem této bakalářské práce bylo zjistit, zda je možné odlišit autory texty na základě toho, jak používají prefixy.

V teoretické části byl stručně popsán obor forenzní lingvistika. Přiblížena byla jeho historie i současnost včetně příkladu sporu o autorství textů podepsaných jménem Josefa Baráka. Dále byla pozornost věnována předmětu zájmu forenzních lingvistů, především tedy anonymním dopisům, jakožto nejvíce zkoumanému typu textu. Následně byly popsány dvě základní metody, které používají forenzní lingvisté – profilování a identifikace autora.

Praktická část byla věnována zodpovězení otázky, zda je používání prefixů u různých autorů odlišné a zda lze identifikovat autora textu podle toho, jak předpony používá. K analýze jsme měli k dispozici dva soubory uměleckých textů (Zrcadla a Trpaslíci), v každém souboru byly převážně texty napsané původním autorem a dva texty napsané jiným autorem.

Při analýze jsme se zaměřili pouze na tři nejfrekventovanější slovní druhy s předponou, tedy substantiva, adjektiva a slovesa. Nejprve jsme se zabývali tím, zda je možné autora odlišit pouze na základě poměru mezi slovy s předponou a slovy bez předpony. Výsledkem u obou souborů bylo konstatování, že podle tohoto poměru autora s jistotou identifikovat nelze, a to u žádného ze zkoumaných slovních druhů.

Dále jsme se pokusili odlišit autora na základě bag-of-words modelů založených na slovech s předponami ve slovníkovém tvaru. Toto kritérium se však neukázalo jako významné při odlišení autorů, protože na jeho základě nebyly podvržené texty dostatečně odděleny od textů původních. K úspěšnému oddělení se sice blížil model založený na slovesech s předponou u souboru Zrcadla, ale ani ten nedokázal podvržené a původní texty oddělit jednoznačně.

Nakonec jsme se autory snažili odlišit na základě bag-of-words modelů založených pouze na předponách jednotlivých slovních druhů. Stejně jako modely založené na celých slovech s předponou nedokázaly autory odlišit ani modely založené pouze na samotných předponách. A to jak u předpon substantiv, tak ani u předpon adjektiv a sloves. Stejně jako u modelů založených na celých slovech zde fungovaly předpony sloves, i zde ale fungovaly pouze částečně, navíc pouze u souboru Trpaslíci.

Výsledek práce byl tedy takový, že se nepodařilo prokázat, že by mezi autory byly v používání předpon takové rozdíly, které by vedly k jednoznačné identifikaci autora. Tato nemožnost odlišit autora je nejspíše zapříčiněna tím, že je autor uměleckého textu do určité míry omezen tématem svého textu a tím, co popisuje.

Seznam použité literatury

JURKA, Michal; FALTÝNEK, Dan, 2017, FORENZNÍ LINGVISTIKA. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy - Nový encyklopedický slovník češtiny*.

KRÁLÍK, Oldřich, 1957. Neruda nebo Barák?, *Literární noviny* 6, s. 6.

MUSILOVÁ, Václava, 2005a. Co je forenzní lingvistika 1. Pojem a možnosti znaleckého zkoumání, předměty zkoumání. *Čeština doma a ve světě* 13, s. 65–70.

MUSILOVÁ, Václava, 2005b. Co je to forenzní lingvistika 2. Metodologické aspekty identifikace autora psaného jazykového projevu v kriminalistice. *Čeština doma a ve světě* 13, s. 97–102.

MUSILOVÁ, Václava, 2006. Co je forenzní lingvistika 3. Otázky lingvistického a psycholingvistického profilování. *Čeština doma a ve světě* 14, s. 137–142.

PLECHÁČ, Petr; FLAIŠMAN, Jiří, 2017. Problém Barák-Neruda z pohledu současné stylometrie. *Česká literatura* 65. s. 743-769.

OWSIANKOVÁ, Hana; FALTÝNEK, Dan; KUČERA, Ondřej, 2018. Genetic analysis of cabbages and related cultivated plants using the bag-of-words model. *Linguistic frontiers* 1, s. 122–132.

SLAVÍČKOVÁ, Eleonora, 1975. *Retrográdní morfematický slovník češtiny s připojenými inventárními slovníky českých morfémů kořenových, prefixálních a sufixálních*. Praha: Academia.

SVARTVIK, Jan, 1968. *The Evans Statements: A Case for Forensic Linguistics*. Göteborg: Elanders boktryckeri aktiebolag.

SVOBODOVÁ, Marie, 1997. Forenzní lingvistika: obsah a možnosti. *Slovo a slovesnost* 58, s. 124-129.

VAŠÁK, Pavel, 1980. *Metody určování autorství*. Praha: Academia.

Seznam obrázků

Seznam příloh