Univerzita Karlova

Filozofická fakulta

Ústav informačních studií a knihovnictví

Diplomová práce

Mgr. Charlotte Panušková

Vzdálené čtení současné české beletrie

Distant Reading of Contemporary Czech Fiction

Praha 2023

Vedoucí práce: Mgr. et Mgr. Čeněk Pýcha, Ph.D.

Tímto bych chtěla poděkovat svému vedoucímu Mgr. et Mgr. Čeňku Pýchovi, Ph.D. za vedení této práce. Můj vděk patří i Ing. Magdě Friedjungové, Ph.D. a Mgr. Josefu Šlerkovi Ph.D. za cenné konzultace.

Prohlášení:

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, dne 17. prosince 2023

Jméno a příjmení

Klíčová slova (česky)

tematické modelování, digitální literární věda, vzdálené čtení, digitální humanitní vědy, současná česká próza

Klíčová slova (anglicky):

topic modelling, digital literary studies, distant reading, digital humanities, contemporary czech fiction

Abstrakt (česky)

Tato práce se zabývá tematickým modelováním současné české prózy pomocí algoritmů LDA a Top2Vec. Zkoumá, jak výsledky tematického modelování korespondují s dosavadními poznatky literární historie. Dále pak analyzuje, jak se tyto výsledky promítají do klasické literární teorie. Práce tak klade důraz na propojení mezi digitálními metodami analýzy textů a klasickými literárněhistorickými a teoretickými pohledy, čímž přináší nový pohled na interpretaci moderních metod v literárním kontextu.

K modelování byl využit veřejně dostupný korpus *Českého národního korpusu*. Korpus byl pro účely práce očištěn a rozdělen do tří subkorpusů podle data prvního vydání děl. Modely algoritmů LDA a Top2Vec byly vytvořeny ze všech tří subkorpusů. Pro výběr nejpřesnějšího modelu práce využívá metriku skóre koherence *Cv*. Výsledky modelů jsou následně porovnány s dosavadními poznatky literární historie. Práce na závěr zdůrazňuje, že tematické modelování představuje spíše aproximaci literárního systému než prostředek k přímému odhalování témat.

Abstract (in English):

This thesis explores the topic modelling of contemporary Czech prose using LDA and Top2Vec algorithms. It examines how the results of topic modelling correspond to existing knowledge in literary history and further analyses how these findings relate to classical literary theory. The study emphasizes the connection between digital methods of text analysis and traditional literary-historical and theoretical perspectives, offering a new interpretation of modern methods within the literary context.

For modelling purposes, the corpus from the *Czech National Corpus* was used. The corpus was cleaned and divided into three subcorpora based on the publication date of the works. Models of both LDA and Top2Vec algorithms were created from all three subcorpora. To select the most accurate model, the thesis employs the coherence score metric *Cv*. The results of the models are then compared with present knowledge in literary history. The conclusion underscores that topic modelling serves as an approximation of the literary system rather than a direct means of revealing themes.

OBSAH

1 Úvod

Tato práce se zabývá vzdáleným čtením současné české prózy. V rámci této práce definuji současnou českou prózu jako prozaická díla vydaná v češtině, která poprvé vyšla po roce 1989. Pro svou analýzu využívám full-texty děl. Veškerá prozaická díla, se kterými pracuji jsem získala z veřejně dostupného korpusu od *Českého národního korpusu.* Jako analytickou metodu pro vzdálené čtení jsem zvolila tematické modelování, v jehož rámci je možné identifikovat specifický počet témat v textech a následně je charakterizovat pomocí seznamů slov.

V oblasti analýzy textu a zpracování přirozeného jazyka se modelování témat stává stále využívanějším přístupem. Na rozdíl od analýzy klíčových slov, která sice poskytuje určitý, ale často nedostatečně hluboký a kontextově omezený náhled na obsah textu, nabízí modelování témat možnost získat komplexnější a jemnější porozumění obsahu. S příchodem této metody vyvstala i otázka, jak ji využít pro analýzu literárních textů. I v této souvislosti vymezuji směřování této práce: jak využít tematické modelování k hlubšímu porozumění (současné české) literatury?

Hlavním aspektem mé práce je zkoumání vzájemného vztahu mezi výsledky (tématy) tematického modelování a dosavadní poznatky literární historie. Vycházím přitom především z kolektivních publikací *V souřadnicích volnosti* a *V souřadnicích mnohosti*. V případě druhé dekády tohoto století, pro kterou zatím souhrn nevyšel, čerpám z kratších statí a článků literárních kritiků.

Nedostatkem blízkého čtení, které je zatím dominantní metodou literární vědy[1], je, že pracuje pouze se zlomkem knih vydaných v daném období.[2] Nezřídka kdy tak přehlíží některé, obzvláště nemainstreamové, proudy literatury. Metody vzdáleného čtení proti tomu

dokáží zpracovat nesrovnatelně větší množství textů v najednou. Konkrétně tematické modelování, se kterým v této práci pracuji, umožňuje odhalit témata a motivy v literárních dílech. Může tak poskytnout hlubší vhled do obsahu a struktury textů, který pak může vést k novým literárním interpretacím a novému porozumění literární tvorby. Je však důležité

[1] Kiene Brillenburg Wurth and Ann Rigney, eds., *The Life of Texts: An Introduction to Literary Studies* (Amsterdam: Amsterdam University Press, 2019), 390.

[2] Franco Moretti, *Grafy, mapy, stromy: Abstraktní modely literární historie*, trans. Olga Čaplyginová (Praha: Karolinum, 2014), 11.

tyto metody vnímat jako podporu a doplněk ke klasickým metodám literární vědy, nikoliv jako jejich nahrazení.

Text mé práce je rozdělen do několika kapitol. Ve druhé kapitole shrnuji současné trendy české literatury a identifikaci jejích hlavních proudů. Třetí kapitola se věnuje některým postupům zakotveným v oboru digital literary studies a definuje pojem vzdálené čtení. Čtvrtá kapitola představuje dvě metody tematického modelování – LDA a Top2Vec. Jejich konkrétní aplikaci popisuji v páté kapitole. Šestá kapitola prezentuje výsledky získané tematickým modelováním. V sedmé kapitole pak tyto výsledky podrobně analyzuji. Osmá kapitola porovnává výsledky tematické analýzy se stanovisky sekundární literatury. V deváté kapitole následně diskutuji dosažené závěry. Celou práci uzavírám desátou kapitolou.

2 Směřování české polistopadové literatury

V této části stručně shrnuji směřování české polistopadové prózy. Ve shrnutí vycházím z kolektivních publikací *V souřadnicích volnosti* a *V souřadnicích mnohosti,* dále pak z článků Evy Klíčové, Aleny Šidákové Fialové a Petra A. Bílka. Nejedná se o vyčerpávající souhrn, nicméně pro účely této práce je dostačující.

Po roce 1989 se český literární trh proměnil k nepoznání. Náhlá liberalizace způsobila, že se do prodeje začaly dostávat knihy, které by za minulého režimu vůbec nevznikly, nevyšly nebo vyšly jen v omezených samizdatových nákladech, případně v exilu. Nové tituly byly nejrůznějších druhů, vydavatelských intencí a úrovně.[3] Už v roce 1991 bylo jasné, že trh je přesycen a sám se neuživí. Registrováno bylo okolo dvou tisíc nakladatelství, navíc výroba, polygrafické služby a papír markantně zdražily, což se odrazilo na ceně knih. Ty přestávaly být atraktivním zbožím.[4] Zhruba v letech 1992–1993 literární prostředí pozbylo soběstačnost a zůstalo odkázáno na různé formy podpory, na kterou je odkázáno více méně dodnes.[5]

Přelom tisíciletí nepřinesl tak zásadní změny jako listopadový převrat, přesto literaturu proměnily vývojové procesy. Díky kvantitativnímu nárůstu knižního trhu a rozšíření internetu počet knih v oběhu strmě narostl.[6] Nástupem internetu se také knižní svět otevřel nadšeným „amatérům“, kteří se začali podílet na tvorbě literatury.[7]

Knižní trh musel reagovat na poptávku čtenářstva, kteří stále méně dychtili po náročnější literatuře, a naopak si žádali stále více „odpočinkovějších“ a „relaxačních“ titulů.[8] Tím se otevřel prostor pro populární, komerčně úspěšnou knižní nabídku. Literatura postupem měnila své sociální postavení. Spisovatel už nebyl vnímán jako společenský mluvčí, který by reflektoval politické dění. Mnoho prozaiků naopak přijalo tezi, že literatura má být apolitická a měla by se soustředit na své specifické, estetické poslání.[9]

[3] Jiří Zizler, ‘Otevřená dekáda’, in *V souřadnicích volnosti: česká literatura devadesátých let dvacátého století v interpretacích*, ed. Petr Hruška (Praha: Academia, 2008), 14.

[4] Zizler, 16.

[5] Zizler, 16.

[6] Karel Piorecký, ‘Česká literární kultura 2001-2010’, in *V souřadnicích mnohosti: česká literatura první dekády jednadvacátého století v souvislostech a interpretacích*, ed. Alena Fialová, Vydání první, Literární řada (Praha: Academia, 2014), 13.

[7] Piorecký, 47.

[8] Zizler, ‘Otevřená dekáda’, 16.

Překvapivě se tak ústředním tématem nestaly (po)listopadové události.[10] Postupem času se však začaly objevovat tituly, které se snažily „o komplexnější reflexi aktuálních společenských událostí“.[11]

Tržní prostředí rozdělilo českou polistopadovou prózu mezi dva póly. Prvním pólem by se dala nazvat „elitní“ literatura, která se držela vyšších uměleckých cílů. Byla čtenářsky náročná, využívala neotřelé postupy a experimentální a esejistické prvky. Na druhém pólu byla literatura, která na více méně rezignovala na myšlenkovou hloubku a poselství, a naopak využívala ustálená schémata a šablony a dávala důraz na akčnost a senzačnost a řídila se komerční podbízivostí. Mezi těmito póly se pohybovala próza, která sice nerezignovala na uměleckou hodnotu, nicméně se snažila oslovit širší okruh čtenářů využitím postupů a prvků populární literatury. Tento nejasně ohraničený okruh se později začal označovat jako „literární mainstream“.[12]

2.1 Devadesátá léta dvacátého století

V devadesátých letech se próza s vyššími uměleckými cíli rozdělila do dvou linií, které však spojoval důraz na autorské „já“, které vnímá a prožívá svět. První linie se snažila o autentičnost, demystifikaci iluzí a vyjádření „holé pravdy“.[13] Byla především vydávána jako autobiografické deníky, memoáry či vzpomínky, které tvořily jeden z nejvýraznějších prozaických trendů.[14] Mezi nejznámější autory a autorky, jejichž díla byla v 90. letech s nálepkou autentičnosti vydávána patří Jan Zábrana, Václav Černý nebo Josef Hiršal s Bohumilou Grögerovou. Deníkových a memoárových forem také v některých svých dílech využíval po listopadu stále populární Bohumil Hrabal.[15]

Druhá linie naopak tíhla k fantasknosti, rozbití příběhu a prolínaní vizí a reálných představ.[16] Problematizovala schopnost poznávat, reprodukovat a hodnotit svět kolem nás. Značná část byla zasazena do města či velkoměsta (především Prahy) s velkým počtem

[9] Lubomír Machala, ‘Próza’, in *V souřadnicích volnosti: česká literatura devadesátých let dvacátého století v interpretacích*, ed. Petr Hruška (Praha: Academia, 2008), 277.

[10] Machala, 278.

[11] Machala, 278.

[12] Alena Fialová, *Česká próza po roce 1989*, Věda kolem nás (Praha: Academia, 2015), 2.

[13] Machala, ‘Próza’, 280.

[14] Machala, 281.

[15] Machala, 282.

[16] Machala, 280.

zákoutí, chodeb a sklepení, které se proměňovaly.[17] Tento proud byl spojován s širším proudem spojovaným se západní vlnou postmoderny. Autoři a autorky, kteří k tomuto proudu hlásili, rozdělují literární historici na několik generací. Mezi nejstarší řadí autory Milana Kunderu a Jana Křesadla, které Machala nazývá „klasici" postmodernismu.[18] Střední generaci zase Aleš Haman nazývá „generací ukradeného příběhu"[19]. Mezi ty se řadí autoři a autorky jako Jiří Kratochvil, Daniela Hodrová, Vladimír Macura, Ivan Matoušek nebo Michal Ajvaz.[20] Nejmladší generaci postmodernistů představují autoři a autorky narození v šedesátých letech. Mezi nejznámější se řadí Jáchym Topol.[21]

Na druhém konci pólu literární spektra se nacházela populární, komerčně úspěšná literatura. Ta nejprve těžila z oživení tradice rodokapsů. Oblíbené se staly i společenské pamflety. Literatura tíhla k senzaci a snaze čtenáře šokovat, proto se často ubíhala k erotickým a sexuálním výjevům. V první polovině devadesátých let byli komerčně úspěšní hlavně autoři a autorky, kteří populárně zaměřenou literaturu psali už za minulého režimu. Mezi takové se řadili hlavně Vladimír Páral a Ludmila Vaňková. V druhé polovině se čtenářský zájem přeorientoval na autory a autorky, kteří debutovali v devadesátých letech. Známou takovou autorkou je například Halina Pawlovská. [22]

Mezi póly náročné a populární četby se nacházely texty, které se pokoušely sloučit postupy obvyklé v populární literatuře s uměleckými prvky. V této šedé zóně se nacházela velká část vydávaných knih. Některé z nich navazovaly na mezinárodní vlnu postmoderny, která se snažila nabídnout literaturu jak běžnému, tak i erudovanému a kultivovanému recipientovi. Čtenářsky úspěšnými se stali autoři Michal Viewegh, Petr Šabach nebo Roman Ludva. O propojení běžné a erudované literatury se také snažil například Pavel Kohout.[23] Historické romány, v českém prostředí tradičně velmi oblíbené, byly v devadesátých letech spíše na ústupu.[24]

---

[17] Machala, 286.

[18] Machala, 287.

[19] Aleš Haman, 'Fikce a imaginace v prózách 90.let', in *Česká próza 90.let 20.století* (České Budějovice, 2002), 25.

[20] Machala, 'Próza', 288–292.

[21] Machala, 293.

[22] Machala, 279.

[23] Machala, 293–296.

[24] Blahoslav Dokoupil, 'Historická próza v roce nula', in *Česká a slovenská literatura dnes* (Praha–Opava: Ústav pro českou literaturu AV ČR, Slezská Univerzita, 1997), 68–70.

2.2 První dekáda jednadvacátého století

Nové milénium na rozdíl od devadesátých let neprošlo tolika turbulentními změnami. Prozaická literatura se vyvíjela bez větších otřesů a nacházela polohu, ve které by se mohla ustálit. Rozdělení na dva póly – „vysokou" a „nízkou" literaturu – přetrvalo. Mezi nimi se pohyboval okruh próz označovaný jako „literární mainstream". Rozdělení se pak odráželo v recenzích kritiků, kteří pak na jednotlivá díla kladli různé nároky. [25]

Další dělicí linií se stala generační příslušnost autorů. Styl a téma děl se více méně odvíjely od toho, kdy autor do literárního pole vstoupil. Do starší generace se řadili autoři, kteří na literární scénu vstoupili již v šedesátých letech. Mezi takové se řadí známí autoři jako Josef Škvorecký, Arnošt Lustig nebo Ota Pavel. Do střední generace se zase řadili autoři narození mezi čtyřicátými až šedesátými léty, kteří začali psát před listopadovou revolucí. K těm se řadili jak Jiří Kratochvil, Daniela Hodrová nebo Michal Ajvaz, ale také Jiří Hájíček, Michal Viewegh nebo Miloš Urban. Mladší a nejmladší generace pak zpravidla do pole literatury vstoupila až po listopadovém převratu. Alena Šidáková Fialová zmiňuje autorky Petru Soukupovou nebo Natálii Kocábovou.[26]

Autenticitní deníková literatura, která po listopadu zažila boom, byla na ústupu. Stejně tak postmoderní experimentální próza už čtenáře tolik nepřekvapovala a neprovokovala. Ústup těchto dvou proudů naopak otevřel prostor pro literaturu zasazenou do konkrétního času a prostoru, reflektující dějinná traumata, život v (post)komunistické společnosti či literaturu reagující na aktuální společenské problémy. [27]

Témata a inspirace autorů se odvíjely v závislosti na jejich věku a historickém kontextu. Starší generace autorů se zaměřovala na politické změny dvacátého století a osobní zkušenosti, zejména život za protektorátu. Generace „střední" byla mnohem různorodější, s některými autory, kteří se inspirovali postmoderní hrou, ironií a experimenty, a dalšími, kteří se zabývali konkrétním životem a pochmurným stavem světa. Mladší generace autorů se zaměřovala na intimní mezilidské vztahy a hledání identity.

Novým trendem v literatuře 21. století byla próza odehrávající se v exotických prostředích, jaké představuje pro středoevropské Čechy například Jižní Amerika nebo Sibiř. Globalizace a otevřenost evropského prostoru také vedly k vydávání próz, které se zabývaly problémy spojenými s evropským kontinentem, ale zároveň zachovávaly

[25] Alena Fialová, 'Próza', in *V souřadnicích mnohosti: česká literatura první dekády jednadvacátého století v souvislostech a interpretacích*, ed. Alena Fialová, Vydání první, Literární řada (Praha: Academia, 2014), 341.

[26] Fialová, 342.

[27] Fialová, 341.

individualizovaný přístup k jednotlivým osudům postav.[28] Prvním takovým titulem bylo dílo Jaroslava Rudiše *Nebe pod Berlínem*, další autorkou, která se v příbězích vydávala za hranice Evropy, byla Petra Hůlová.[29] Podle Aleny Šidákové Fialové se však „[n]ejvýraznějším trendem […]stal tematický obrat k období čtyřicátých až osmdesátých let dvacátého století".[30] Autoři a autorky, kteří se na takovou tvorbu zaměřovali byli např. Václav Chochola, Irena Dousková nebo Petr Šabach.[31]

2.3 Druhá dekáda jednadvacátého století

Velký souhrn české prozaické literatury jednadvacátého století zatím nevyšel, proto jsem vycházela z vícero článků, debat a kratších komentářů, které směřování současné české prózy shrnují a komentují. Jedna taková debata se odehrála na půdě Ústavu pro českou literaturu AV ČR s názvem *První bilance*. Souhrn debaty přinesla v časopisu *Host* literární kritička Eva Klíčová, která se debaty sama zúčastnila. Účastníci debaty se shodli na tom, že české literatuře dominují romány, které jsou zasazeny do konkrétního místa a času. Moderátor debaty, literární vědec Martin Lukáš též poznamenal, že se oblibě kritiků i čtenářů poslední roky těší příběhově založené knihy, které jsou stylisticky jednodušší a vypráví jeden příběh.[32]

Dalším rysem literatury druhé dekády jsou historické romány. Podle již zmíněné Evy Klíčové je pro českou literaturu po roce 2000 nejpříznačnější psaní o minulosti, ať už si autor minulost pamatuje, či nikoliv. Romány jsou převážně zasazené do 40. až 80. let, a ze své pozice aktualizují dějiny. Též pro ně není neobvyklé „postmoderní" mísení naznačených postupů. Mezi nejznámější autorky, které se ve své tvorbě vrací do tohoto období jsou Kateřina Tučková a Alena Mornštajnová. Z autorů je možné zmínit Antonína Bajaju nebo Jaroslava Kovandu. Knihy zasazené do staršího období naopak nabízí romantický eskapismus.[33]

---

[28] Fialová, 342.

[29] Fialová, 351–354.

[30] Fialová, 343.

[31] Fialová, 355–358.

[32] Eva Klíčová, 'Kritika v osamění', Accessed 3 November 2023, 25 January 2020, https://www.h7o.cz/clanky/12683-kritika-v-osameni.

[33] Eva Klíčová, 'Historický román v současné české literatuře versus téma dějin v české próze (po roce 2000)', Accessed 3 November 2023, 11 January 2019, https://www.czechlit.cz/cz/feature/historicky-roman-v-soucasne-ceske-literature-versus-tema-dejin-v-ceske-proze-po-roce-2000/.

S tímto proudem také souvisí obliba biografických románů v pravém slova smyslu, ve kterých je patrná práce s deníky a s dalšími zdroji.[34] Podle Aleny Šidákové Fialové lze v desátých letech 21. století pozorovat výrazný fenomén próz, které jsou inspirované reálnými životními osudy. Jde o knihy, které se pohybují na hranici literatury umělecké a faktografické a většinou přinášejí životopisné příběhy jedné nebo více postav. Příznačná je pro tuto prózu také důkladná práce s dobovými prameny.[35] Biografické romány psali například Jan Němec, Irena Dousková nebo Martin Reiner, který za svůj román *Básník* získal cenu Magnesia Litera.

Literární teoretik Petr A. Bílek je toho názoru, že „[j]ádrem současné české prózy jsou, tak jako vždy a všude, krizové vztahy.“[36] V textu pro deník *Aktuálně.cz* zmiňuje několik autorek, které se ve své tvorbě na tyto „krizové vztahy“ zaměřují. Především vyzdvihuje tvorbu Radky Denemarkové, dále pak Petry Soukupové, Anny Bolavé, Viktorie Hanišové nebo Jakuby Katalpy. Podle Bílka je literatura druhé dekády jednadvacátého století zavalena balastem a z knih většinou vznikne „jakási amorfní hmota“[37]. Výjimku podle Bílka tvoří tvorba Jáchyma Topola a Karla Sidona.

[34] Alena Šidáková Fialová, ‘Biografický román v současné české literatuře’, Accessed 1 November 2023, 7 November 2018, https://www.czechlit.cz/cz/feature/biograficky-roman-v-soucasne-ceske-literature/.

[35] Alena Šidáková Fialová, ‘Česká próza po roce 2000’, Accessed 1 November 2023, 1 September 2015, https://www.czechlit.cz/cz/feature/ceska-proza-po-roce-2000/.

[36] Petr A. Bílek, ‘Dekáda v české literatuře: Hodně balastu a málo velkých románů, vyniká Topol či Sidon’, Accessed 30 October 2023, 27 December 2019, https://magazin.aktualne.cz/kultura/literatura/ceska-literatura-2010-2019/r~eacde10e28ca11ea8776ac1f6b220ee8/.

[37] Bílek.

3 Digital literary studies

Digital literary studies, nebo také computational literary studies, představují přístup, který spadá do souboru metod multivědního oboru digital humanities. Ten se objevil s příchodem digitálních technologií. Jeho základem je využití digitálních či komputačních nástrojů, které mechanicky zpracovávají objekty humanitního zkoumání.[38] V případě digital literary studies se většinou jedná o literární texty nebo bibliografická metadata, jejichž součástí můžou být informace o autorech, žánrech knih, vydáních, překladech atp. Pod deštník přístupu digital literary studies se schová velké množství metod, jak data zpracovávat. V základu lze tyto metody rozdělit podle toho, zda zkoumaná data sestávají z plných textů (full-textů) či metadat. V případě analýzy plných textů má přístup digital literary studies často blízko k metodám korpusové lingvistiky.[39] Výzkumy využívají metody jako klíčová analýza slov[40] či frazeologie[41].[42] Při práci s metadaty se používá např. síťová analýza.[43]

Literární vědci Ray Siemens a Susan Schreibman zasazují počátky digital literary studies do poloviny minulého století, kdy vznikl projekt Roberta Busa *Index Thomisticus*,[44] v jehož rámci zdigitalizoval během 49 let texty Tomáše Akvinského. V druhé polovině minulého století se ale objevilo i několik dalších prací, které by se daly označit za milníky v oblasti digital literary studies.[45]

---

[38] Patrik Svensson, 'Beyond the Big Tent', in *Debates in the Digital Humanities*, ed. Matthew K. Gold (University of Minnesota Press, 2012), 67–71.

[39] David L. Hoover, Jonathan Culpeper, and Kieran O'Halloran, eds., 'Introduction', in *Digital Literary Studies*, 0 ed. (Routledge, 2014), 2.

[40] Jonathan Culpeper, 'Keywords and Characterization: An Analysis of Six Characters in Romeo and Juliet', in *Digital Literary Studies*, ed. David L. Hoover and Kieran O'Halloran ed. (Routledge, 2014); Michel JB, Shen YK, Aiden AP, Veres A, Gray MK; Google Books Team, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, Aiden EL, 'Quantitative Analysis of Culture Using Millions of Digitized Books', 2011, 176–182.

[41] Iva Novakova and Dirk Siepmann, eds., *Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives* (Springer Nature, 2019).

[42] Hoover, Culpeper, and O'Halloran, 'Introduction', 2.

[43] Jeffrey Drouin, 'Close- and Distant-Reading Modernism', *The Journal of Modern Periodical Studies* 5, no. 1 (2014): 110–135.

[44] Ray Siemens and Susan Schreibman, eds., 'Editors' Introduction', in *A Companion to Digital Literary Studies*, 2007, xvii.

3.1 Vzdálené a blízké čtení

V současném tisíciletí byla velkým milníkem kniha literárního historika Franca Morettiho *Grafy, mapy, stromy* z roku 2005. V ní poprvé použil a rozvedl pojem vzdálené či distanční čtení (*distant reading*).[46] Vzdálené čtení zde definuje jako „přístup, pro nějž vzdálenost neznamená překážku, nýbrž specifickou formu poznání“,[47] a pojem vymezuje vůči tzv. metodě blízkého čtení (*close reading).*[48] Ta podle něj striktně vychází pouze z textu samotného.[49] Metody by se tak daly přirovnat k přístupu *kvantitativnímu* (*distant reading*) a *kvalitativnímu* (*close reading*). Myšlenku vzdáleného čtení Moretti později rozvinul v knize *Distant reading*.[50] V knize *Grafy, mapy, stromy* Moretti pracuje s metadaty k literárním textům, která následně převádí do podoby vizuálního zpracování. Morettiho základní myšlenkou bylo vytvoření systematického pohledu na literární produkci a hledání obecnějších trendů převážně skrz vizualizace. Dalším důvodem, proč Moretti tento přístup tak propagoval, byla nezávislost na „oficiálním kánonu“, na který se velká část tradičních literárních historiků odkazuje.[51]

Na Morrettiho práci navázal další literární vědec a průkopník přístupu digital literary studies Matthew Jockers. Svůj přístup ke zpracování literatury představil v roce 2013 ve své knize *Macroanalysis: Digital Methods and Literary History.* V ní se zabývá jak zpracováním metadat, tak textovou analýzou literárních textů zahrnující stylometrii a tematické modelování. Jockers rozvíjí Morettiho vzdálené čtení a nahrazuje ho pojmem „makroanalýza“. V knize představuje konkrétní statistické nástroje a metody strojového učení k identifikaci vzorů v textu. Metody, které Jockers využívá pro zpracování literatury, pak autor spolu s Rosamond Thalken sepsal v knize *Text Analysis with R: For Students of Literature*.[52]

---

[45] Rosanne G. Potter, *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric* (University of Pennsylvania Press, 1989); Richard J. Finneran, *The Literary Text in the Digital Age* (University of Michigan Press, 1996).

[46] Franco Moretti, *Distant Reading* (London: Verso, 2013).

[47] Moretti, *Grafy, mapy, stromy: Abstraktní modely literární historie*, 7.

[48] Moretti, 8.

[49] Moretti, 7.

[50] Moretti, *Distant Reading*.

[51] Moretti, *Grafy, mapy, stromy: Abstraktní modely literární historie*, 11.

[52] Matthew L. Jockers and Rosamond Thalken, *Text Analysis with R: For Students of Literature*, Quantitative Methods in the Humanities and Social Sciences (Cham: Springer International Publishing, 2020).

Práce Franca Morettiho a Matthewa Jockerse byla nejprve vědeckou veřejností oslavována,[53] později se však ze strany vědecké obce začala ozývat kritika. Nejznámější kritičkou Morettiho a Jockerse je literární teoretička Katherine Bode, která svoji knihu *A World of Fiction: Digital Collections and the Future of Literary History* otevírá kapitolou kritickou vůči dvěma zmíněným vědcům. Bode Morettimu s Jockersem vyčítá, že svá zjištění předkládají jako objektivní, stabilní a samozřejmá, zatímco oba zkoumají jen omezenou část literatury na vybraném literárním vzorku, který je i tak zatížen svým biasem. Dále oba kritizuje za netransparentnost jejich metod, které sice popisují, ale své konkrétní kódy a datasety nesdílejí.[54]

3.2 Digital literary studies v Česku

Digital literary studies se v českém prostředí začaly objevovat později než v anglofonním a popularitě se začaly těšit až v posledních několika letech. Na vině je několik faktorů, mezi něž patří nedostatek (veřejně i neveřejně) dostupných dat nebo nedostatečný software přizpůsobený českému jazyku. Většina výzkumu v oblasti digital literary studies (a digital humanities obecně) totiž stojí na datech, která jsou těžko dostupná. Výjimkou je výzkum nad *Korpusem českého verše*, který probíhá na Ústavu pro českou literaturu AV ČR.[55]

Jediný obsáhlejší a momentálně dostupný beletristický literární korpus je zprostředkovaný *Českým národním korpusem,* akademickým projektem Ústavu Českého národního korpusu.[56] Ten sice obsahuje především publicistická díla, nicméně obsahuje i díla literární, z nichž největší část tvoří díla vydaná po roce 1990. *Český národní korpus* ke svému korpusu vydal několik rozhraní, která práci s ním usnadňují.

Svůj kvantitativně-korpusový beletristický výzkum na korpusu *Českého národního korpusu* provedl Richard Změlík, který se zabýval výskytem barev v několika korpusech, subkorpusech a dílech konkrétních autorů.[57] Referenčním korpusem mu byl reprezentativní

[53] Ted Underwood, 'A Dataset for Distant-Reading Literature in English, 1700-1922', 7 August 2015, https://tedunderwood.com/2015/08/07/a-dataset-for-distant-reading-literature-in-english-1700-1922/.

[54] 'Abstraction, Singularity, Textuality The Equivalence of "Close" and "Distant" Reading', in *A World of Fiction: Digital Collections and the Future of Literary History*, by Katherine Bode (University of Michigan Press, 2019), 17–35.

[55] Petr Plecháč et al., 'Korpus českého verše', accessed 19 August 2023, https://versologie.cz/v2/web_content/corpus.php.

[56] Michal Křen et al., 'SYN v9: Large Corpus of Written Czech', 2021, http://hdl.handle.net/11234/1-4635.

[57] Richard Změlík, *Konceptualizace barev v narativní fikci na pozadí kvantitativních modelů* (Olomouc: Univerzita Palackého v Olomouci, 2019).

synchronní korpus SYN 15. Dále jevy sledoval v subkorpusu SYN FIC, sestávajícího ze subkorpusu české prózy SYN FIC NOV a české poezie SYN FIC VER. Nakonec využil několik vybraných autorských korpusů próz.[58] Výskyty barev následně porovnal a interpretoval. Jako barvu Změlík bral nejen adjektiva (např. červený), která jsou v jazyce zastoupena nejvíce, ale i „substantivum pojmenovávající barvu jakožto substanci (červeň), verbum označující procesualitu (červenat) a adverbium jako příznak děje, stavu, situace (červeně).“[59] Barva pak je frekvenční suma těchto čtyř slovních druhů.

Celkem nepřekvapivě mezi nejčastějšími barvami byla bílá a černá, za kterými následovaly chromatické barvy červená, modrá a další.[60] Nejnižší míru divergence od referenčního korpusu vykazoval korpus SYN FIC NOV, naopak v korpusu SYN FIC VER se častěji vůči referenčnímu korpusu vyskytovaly achromatické barvy černá a bílá. Autorské subkorpusy pak vykazovaly nejvyšší míru divergence od korpusu SYN 15.[61]

V českém prostředí je také významná publikace *Kvantitativní analýza žánrů* Miroslava Kubáta, ve které autor za použití stylometrických metod analyzoval rozdílnost žánrů díla. Jako metody využil analýzu slovního bohatství, tematickou koncentraci textu, vzdálenost sloves, průměrnou délku tokenu, aktivitu a deskriptivitu, n-gramy, a nakonec nejfrekventovanější slova.[62] Ve svém výzkumu se omezil pouze na díla Karla Čapka, čímž vyfiltroval nežádoucí jazykové styly jiných autorů. Jeho rozhodnutí sice vylučuje obecnost, nicméně dobře ilustruje všechny použité klasifikační metody a jejich nedostatky. Jeho korpus sestával z 33 Čapkových děl, které rozdělil do osmi žánrů – studie, román, cestopis, sloupek, pohádka, dopis, báseň a povídka. Všechny metody byly schopny žánry obstojně rozeznat s tím, že čím složitější metoda byla, tím byla přesnější. Nicméně s obtížností metody klesala možnost lingvistické interpretace. Překvapivým závěrem bylo, že báseň byla druhým „nejméně specifickým“ žánrem, tedy stylistické metody měly největší potíž báseň rozeznat od ostatních žánrů.[63]

Co se týče poezie, versologický tým Ústavu pro českou literaturu AV ČR zprostředkovává *Korpus českého verše,* „lemmatizovaný, foneticky, morfologicky,

[58] Změlík, 115.

[59] Změlík, 122.

[60] Změlík, 122–123.

[61] Změlík, 123.

[62] Kubát, Miroslav, *Kvantitativní analýza žánrů* (Ostrava: Ostravská univerzita, Filozofická fakulta, 2016), 8.

[63] Kubát, Miroslav, 107.

metricky a stroficky anotovaný korpus české poezie 19. a počátku 20. století.“[64] Korpus sestává z 1 689 básnických sbírek a obsahuje 76 669 básní.[65] Od roku 2021 je na něj navázaný vědecký projekt s názvem „Analýza motivických klastrů z oblasti aktuálních kulturně-společenských témat a jejich aplikace na materiál uměleckých textů 19. a počátku 20. století“, který si klade za cíl zanalyzovat možnost vytvoření tematických shluků z textů *Korpusu českého verše*.[66]

[64] Plecháč et al., ‘Korpus českého verše’.

[65] Plecháč et al.

[66] Daniela Iwashita et al., ‘Analýza motivických klastrů z oblasti aktuálních kulturně-společenských témat a jejich aplikace na materiál uměleckých textů 19. a počátku 20. století’, 2021, https://ucl.cas.cz/projekt/analyza-motivickych-klastru/.

4 Tematické modelování

Nadcházející kapitola se zaměří na konkrétní metodu digital literary studies – tematické modelování a přístupy k analýze témat v textech. Představí základní koncepty spojené s pojmem téma a následně popíše metody tematického modelování.

4.1 Co je to téma

Tato práce pracuje s přístupem zvaným *topic modelling*, do češtiny přeložený jako *tematické modelování*. Hned v počátku však narazíme na problém překladu. České slovo *téma* totiž odpovídá nejen anglickému slovu *topic*, ale i *theme*. Podle výkladového slovníku je *topic* „a subject that is discussed, written about, or studied“[67], tedy námět nebo předmět, o kterém se diskutuje, píše nebo který se studuje. *Theme* je pak „topic of discussion or writing“[68], ale taky „the main idea of a work of literature or art“[69]. *Theme* se tedy dá chápat jako *topic*, ale i jako implicitní či abstraktní myšlenka díla, která nemusí souviset s *topic*. V češtině tyto dva pojmy splývají, jelikož jsou oba vyjádřeny slovem *téma*.

S pojmem *téma* též hojně pracuje i literární věda, která tématem obvykle myslí hlavní myšlenku, smysl či ideu díla, kterou autor v díle rozvíjel.[70] Historicky je téma synonymně používáno s pojmy námět, látka, či motiv, ale i s pojmy jako obsah, forma nebo topos.[71] Každý směr literární vědy k tématu přistupoval a chápal je rozdílně, a téma tak nacházel rozdílnými způsoby. Např. Hodrová za téma považuje „obecnější motivické komplexy a toposy, které se vracejí v různých dílech […] přesahují jejich významovou strukturu a nezřídka mají filozofický a ontologický význam […] případně se vztahují k archetypům […] a nejsou významné jen v jedné epoše, ale v rámci celé lidské kultury“.[72]

---

[67] Cambridge University Press, ‘Topic’, in *Cambridge Academic Content Dictionary* (Cambrige University Press, 2009), https://dictionary.cambridge.org/dictionary/english/topic.

[68] Cambridge University Press, ‘Theme’, in *Cambridge Academic Content Dictionary* (Cambrige University Press, 2009), https://dictionary.cambridge.org/dictionary/english/theme.

[69] Cambridge University Press.

[70] Ondřej Sládek and et. al., eds., ‘Téma’, in *Slovník literárněvědného strukturalismu* (Praha, Brno: Host, 2018), 751.

[71] Sládek and et. al., 751.

[72] Daniela Hodrová and et. al., ... *na okraji chaosu...: Poetika literárního díla 20. století* (Praha: Torst, 2001), 739.

Tematologií se prvně a nejvíce zabývali formalisté, pak strukturalisté a literární vědci hlásající se ke směru „Nová kritika“.[73] Jak strukturalisté, tak noví kritici vycházeli z poznatků ruského formalismu. Téma je pro tyto směry plně determinované strukturou a hierarchií literárního díla a skládá se podle nich z motivů.[74] Motiv je podle nich nejmenší a nedělitelnou jednotkou textu a zároveň základem tematické výstavby. Stejný motiv může být vyjádřen různými textovými vzorci, a naopak stejnými textovými vzorci mohou být vyjádřeny různé motivy.[75] Podle lingvistů a literárních vědců Alexandera Veselovského a Boris Tomaševského se motivy sdružují ve složitější celky, které dohromady dávají syžety (celkové výstavby děje). Tomaševskij přímo říká, že „[m]otivy se navzájem spojují, a tvoří tak tematické spoje díla. Z tohoto hlediska je fabule souhrnem motivů v jejich logických, příčinných a časových souvislostech, syžetem je souhrn těchže motivů v následnosti a těch souvislostech, v jakých jsou předvedeny v díle“.[76]

Literární vědec a komparatista Claudio Guillén o několik desítek let později zmiňuje, že tematologický výzkum je spjat nejen s výzkumem morfologickým, se kterým jej spojovali formalisté a strukturalisté, ale i genologickým, tedy s výzkumem literárních žánrů. [77] Ve své knize pak předkládá šest hledisek, jak žánry analyzovat. Mezi ně přidal i hledisko strukturální. Žánr podle tohoto hlediska patří do určitého souboru možností, alternativ a vzájemných vztahů. Každý žánr naplňuje určitou funkci v celkovém literárním systému, jedné velké, složitě uspořádané a organické jednotce.[78]

Matthew Jockers, který se jako první zabýval tematickým modelováním literárních textů, distinkci pojmů *topic*, *theme* a *motiv* ve své definici maže. Podle něj je téma typ literárního materiálu, který se s nějakou frekvencí objevuje napříč literárními texty (korpusem).[79] Díky smazání těchto terminologických hranic tak může téma definovat

[73] Ondřej Sládek and et. al., eds., ‘Motiv’, in *Slovník literárněvědného strukturalismu* (Praha, Brno: Host, 2018), 472.

[74] Sládek and et. al., ‘Téma’, 751.

[75] Sládek and et. al., ‘Motiv’, 472.

[76] Boris Viktorovič Tomaševskij, *Teorie literatury*, trans. Renáta Štindlová and Karel Štindl (Praha: Lidové nakladatelství, 1970), 128.

[77] Anna Housková, Alexandra Berendová, and Mariana Housková, trans., ‘Témata: Tematologie’, in *Mezi jedotou a růzností: Úvod do srovnávací literární vědy*, by Claudio Guillén (Praha: Triáda, 2008), 205.

[78] Anna Housková, Alexandra Berendová, and Mariana Housková, trans., ‘Literární žánry: Genologie’, in *Mezi jedotou a růzností: Úvod do srovnávací literární vědy*, by Claudio Guillén (Praha: Triáda, 2008), 120.

[79] Matthew Lee Jockers, *Macroanalysis: Digital Methods and Literary History*, Topics in the Digital Humanities (Urbana: University of Illinois Press, 2013), 123.

několika klíčovými slovy, která jsou výstupem algoritmů tematického modelování. Jockers se pojetím konceptu *téma* inspiroval již zmíněným Alexandrem Veselovským a jeho bratrem Aleksejem. Ti ve svých dílech mimo jiné komparativní metodou hledali tematické vlivy západu v dílech ruských autorů.[80]

4.2 Popis metod

Modelování témat je metoda, která je součástí přístupu zpracování přirozeného jazyka (NLP). Používá se k nalezení témat v případě, kdy máme velké množství textů. Primární cíl je automaticky identifikovat skrytá (latentní) témata v textech, o kterých nemáme předchozí znalost. Existuje několik přístupů k modelování témat, z nichž každý má své výhody a omezení. Mezi nejběžněji používané patří Latent Dirichlet Allocation (LDA), což je statistický model, který se snaží odhalit skrytá témata v textových datech, a Latentní Sémantická Analýza (LSA), která analyzuje významy a vztahy mezi slovy v textu. Tyto metody umožňují lépe porozumět obsahu textu a nalézt v nich skryté vzory a témata. Kromě LDA a LSA existuje také Non-Negative Matrix Factorization (NMF). V poslední době se objevují i modernější přístupy, jako jsou BERT-topic a Top2Vec, které využívají pokročilé modely strojového učení pro ještě přesnější a efektivnější modelování témat. Metody LDA a Top2Vec teď důkladněji popíšu.

4.3 LDA

Pravděpodobnostní model LDA představila v roce 2003 trojice vědců David Blei, Andrew Ng a Michael Jordan.[81] Od té doby se drží na pozici jednoho z nejvýznamnějších přístupů k tematickému modelování. Základním kamenem LDA jsou dva klíčové předpoklady. Zaprvé, model předpokládá, že každý dokument, ať už je to článek, textový soubor nebo jiný textový záznam, je vytvořen kombinací různých témat. Tento předpoklad reflektuje skutečnost, že texty v reálném světě často zahrnují různé aspekty a témata. Zadruhé, LDA předpokládá, že každé téma je tvořeno korpusem slov, která jsou navzájem podobná významem nebo kontextem.[82]

Slova jsou algoritmem sdružena do tematických shluků na základě pravděpodobnosti svého výskytu v dokumentech. Ta, která se často společně vyskytují v

[80] Jockers, 119.

[81] David M Blei, 'Latent Dirichlet Allocation', 2003.

[82] Blei, 993.

různých dokumentech, jsou pravděpodobně spojena s tím samým tématem. Tímto způsobem LDA vytváří shluky slov, které tvoří jednotlivá témata. Tento přístup je v podstatě obdobou metody „Bag of Words“, ale s přidanou vrstvou pravděpodobnostního modelování, která umožňuje lépe zachytit strukturu textu. Následně je na vědci, vědkyni či na osobě provádějící analýzu, aby tyto shluky pojmenoval.

Algoritmus má několik parametrů, které lze upravovat. Jedním z klíčových parametrů je počet tematických shluků, které chce uživatel namodelovat. Tento parametr ovlivňuje, kolik různých tematických kategorií model v datech identifikuje. Nesprávně zvolený počet témat může vést k nedostatečnému nebo nadměrnému zdůraznění určitých témat. V praxi se počet témat vybírá buď na základě znalosti korpusu, jako to udělal např. Radim Hladík ve svém sociologickém výzkumu[83], nebo se natrénuje několik modelů s různým počtem témat. [84] Pomocí dalších metrik (jako např. topic coherence nebo topic diversity) se vybere nejpřesnější model.

Dalším důležitým parametrem je „alpha“, který ovlivňuje distribuci témat v jednotlivých dokumentech. Vyšší hodnoty alfa mohou způsobit, že jednotlivé dokumenty budou obsahovat širší spektrum témat, zatímco nižší hodnoty alfa vedou k tomu, že dokumenty budou více specializované na určitá témata. „Beta“ je dalším významným parametrem, který řídí distribuci slov v jednotlivých tématech. Vyšší hodnoty beta znamenají, že témata budou obsahovat více různých slov, zatímco nižší hodnoty beta způsobí, že témata budou obsahovat méně, ale častěji se opakujících slov.[85]

4.4 Top2Vec

Model Top2Vec představil v roce 2020 Dimo Angelov.[86] Je tedy novější než LDA a je také založen na odlišném principu. Nejedná se totiž o pravděpodobnostní model, ale o model založený na neuronových sítích. Nerozřazuje předem určený počet témat mezi dokumenty,

[83] Radim Hladík, ‘Modelování témat v české sociologii: typy autorství a citační ohlas v odborných textech’, n.d., 159–190.

[84] Mats Dahllöf and Karl Berglund, ‘Faces, Fights, and Families: Topic Modeling and Gendered Themes in Two Corpora of Swedish Prose Fiction’, 2019; Anna Tesaříková, ‘Topic Modeling for Corpus of Czech Verse’ (bakalářská práce, Praha, České vysoké učení technické, 2022); Martin Bendík, ‘Automatic Detection of Topics in Poetic Texts’ (diplomová práce, Praha, České vysoké učení technické, 2023).

[85] Annibale Panichella, ‘A Systematic Comparison of Search-Based Approaches for LDA Hyperparameter Tuning’, *Information and Software Technology* 130 (February 2021): 4.

[86] Dimo Angelov, ‘Top2Vec: Distributed Representations of Topics’ (arXiv, 2020), http://arxiv.org/abs/2008.09470.

ale přiřadí jednotlivým dokumentům jedno téma. Jejich počet navíc neurčuje uživatel, nýbrž jsou vypočteny z algoritmu. Od LDA ho pak odlišuje i schopnost zachytit kontext a synonyma.

Model nejprve pomocí algoritmu doc2vec dokumenty zanese do vektorového prostoru, kde vytvoří distribuovanou reprezentaci slov. Tím zachytí syntaktické a sémantické vztahy mezi slovy na základě jejich kontextu.[87] Následně pak zredukuje dimenzi prostoru pomocí algoritmu Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP). Původní rozměr prostoru se běžně pohybuje okolo 300 dimenzí, redukovaný je na obvykle 5 dimenzí.[88] Z redukovaného prostoru pak vytvoří shluky dokumentů pomocí algoritmu Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN).[89] Shluky jsou tvořeny dokumenty, které se pak podílejí na tvorbě slov tématu. Každý dokument je tak přiřazen k jednomu tématu – shluku. V původním 300dimenzionálním prostoru jsou pak pomocí kosinové vzdálenosti spočteny centroidy – vektory slov, které definují dané téma. Je důležité zmínit, že vektorový prostor sestává jak z dokumentů, tak z témat. Dokumentům pak lze počítat vzdálenost pomocí kosinové vzdálenosti, která nabývá hodnot od 0 do 1, kde 0 znamená nejvyšší vzdálenost a 1 nejnižší.

Podle posledních výsledků jsou modely založené na neuronových sítích přesnější než modely pravděpodobnostní.[90]

4.5 Skóre koherence

Jelikož jsou tematické modely založené na různých předpokladech, tudíž i jejich závěry jsou vyjádřeny jinou metrikou, je potřeba mít externí metriku, která výsledky porovná. Nejčastější způsob, jak se tematické modely porovnávají, je tzv. skóre koherence. Tato metrika obecně vyjadřuje, jak interpretovatelná a koherentní jsou výsledná témata modelů. Slova v tématech by se tedy měla objevovat v podobném kontextu, či by si měla být sémanticky podobná.[91]

---

[87] Angelov, 2.

[88] Angelov, 6.

[89] Angelov, 6.

[90] Ferhat D Zengul et al., 'A Practical and Empirical Comparison of Three Topic Modeling Methods Using a COVID-19 Corpus: LSA, LDA, and Top2Vec', 2023; Roman Egger and Joanne Yu, 'A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts', *Frontiers in Sociology* 7 (6 May 2022).

[91] Bendík, 'Automatic Detection of Topics in Poetic Texts', 10.

K určení metriky se používá několik způsobů, podle kterých se metrika vypočítává. Mezi oblíbené patří C_{UCI}, C_{UMass} a C_V. Zatímco metriky C_{UCI} a C_{UMass} jsou založené na kookurenci (spoluvýskytu) a frekvenci slov v korpusu, metrika C_V měří sémantickou podobnost slov. Podle vědeckého kolektivu vedeného Röderem, který metriky porovnával, koreluje metrika C_V nejvíce s hodnocením témat lidskými respondenty.[92] Metrika C_V leží na škále od 0 do 1, kde vyšší číslo znamená vyšší koherenci a interpretovatelnost. Je nutné podotknout, že metrika neměří přesnost modelu.

4.6 Literature review

Jedni z prvních průkopníků tematického modelování byli již zmíněný Matthew Jockers a David Mimno, kteří se ve svém výzkumu zaměřili na tematické modelování anglicky psané literatury z 19. století.[93] Statistickým testem (bootstrappingem) pak zkoumali, jaká témata spojená s autorkami a jaká s autory. Jejich korpus sestával z 3 346 děl, z čehož 1 364 bylo napsáno autorkami, 1 770 autory, a u 145 nebylo možné gender autorské entity zjistit. Korpus zbavili stop-slov, názvů a jmen. Pro svůj výzkum využili implementaci algoritmu LDA v nástroji Mallet, kde extrahovali 500 témat. Na trénování modelu použili pouze podstatná jména. Knihy rozdělili do cca 1000slovných celků.[94]

Prvním závěrem bylo, že některá témata byla více přítomna mezi autorkami a některá mezi autory. Pro autorky to bylo téma „Female fashion“, pro autory téma „Enemies“.[95] Následně pak testovali, nakolik lze klasifikátorem určit, zda byla kniha napsána autorem či autorkou. Vstupy klasifikátoru byly témata přiřazená algoritmem LDA. Senzitivita klasifikátoru byla 81 %, tedy ve více než 8 z 10 děl byl gender přiřazen správně.

Podobný výzkum provedla dvojice vědců Mats Dahllöf a Karl Berglund na dvou korpusech švédské literatury.[96] První korpus zahrnoval klasická švédská literární díla

[92] Michael Röder, Andreas Both, and Alexander Hinneburg, ‘Exploring the Space of Topic Coherence Measures’, in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (WSDM 2015: Eighth ACM International Conference on Web Search and Data Mining, Shanghai China: ACM, 2015), 399–408.

[93] Matthew L. Jockers and David Mimno, ‘Significant Themes in 19th-Century Literature’, *Poetics* 41, no. 6 (December 2013): 750–769.

[94] Jockers and Mimno, 754.

[95] Jockers and Mimno, 758–760.

[96] Dahllöf and Berglund, ‘Faces, Fights, and Families: Topic Modeling and Gendered Themes in Two Corpora of Swedish Prose Fiction’.

napsaná v období 1821 až 1941, zatímco druhý korpus obsahoval bestsellery publikované v letech 2004 až 2017. I Dahllöf a Berglund spojovali témata s genderem, nicméně na rozdíl od Jockerse a Mimna zkoumali, jak jsou tato témata spojena s genderem postav.

Stejně jako Jockers a Mimno využili pro tematické modelování implementaci algoritmu LDA v nástroji Mallet.[97] Do analýzy zahrnuli pouze podstatná jména a slovesa. Během průběhu studie zkoumali možnost extrakce 20, 40 a 100 témat, přičemž každé téma bylo charakterizováno 12 nebo 24 klíčovými slovy. V korpusu klasické literatury byla jako „nejvíce ženská" témata identifikována tato: „Dance, music, entertainment", „Family relations" a „Mental states, existential reflection". Jako „mužská" pak „Authority", „God, religion, faith" a „Money, work, trade".[98] Témata obecně reflektující tehdejší genderové role a stereotypy, tedy že ženy se více staraly o „rodinný krb", kdežto muži více vládli a vydělávali peníze. Podobné tendence nalezli i v korpusu švédských bestsellerů.[99]

Tým vědkyň Tatiany Sherstinové si ke svému zkoumání vybraly korpus ruských povídek z počátku 20. století. Automatické anotování témat pak porovnaly s anotacemi od literárních vědců. Jejich korpus sestával z 310 povídek, které byly napsány mezi lety 1900–1930. Korpus rozdělily na tři časová období: 1. období před 1. světovou válkou 1900–1913; 2. období 1. světové války a ruské občanské války 1914–1922; 3. období porevoluční éry let 1923–1930. Po počátečních experimentech si vědkyně zvolily jako metodu NMF z knihovny gensim. Jedno téma reprezentovaly 10 slovy, témat hledaly od 5 do 20.

Jejich výsledky ukázaly, že v jednotlivých historických obdobích převažují různé tematické třídy. Například v období raného 20. století byla nejvíce zastoupena témata spojená s popisem lidí, profesemi, každodenními realitami a abstraktními pojmy. Ve druhém období (první světová válka a revoluce) to byla témata spojena s válkou, vírou a přírodou. Ve třetím období (raný sovětský režim) se objevila témata týkající se práce v továrnách, vesnického života a lovů. Automaticky vygenerovaná témata pak porovnaly s anotacemi literárních vědců. U některých témat byla vyšší korelace mezi klíčovými slovy a tématy určenými automaticky a těmi manuálně anotovanými odborníky, u jiných témat tato korelace chyběla zcela. Nejlepší shoda byla u témat „War", „Nature", „Family" a „Religion".[100]

[97] Dahllöf and Berglund, 97.

[98] Dahllöf and Berglund, 102.

[99] Dahllöf and Berglund, 103.

Tematickým modelováním se zabývaly i vědkyně Inna Uglanova a Evelyn Gius. Ve svém výzkumu se zaměřily na metriku skóre koherence *Cv*, která počítá, jak koherentní témata modelů jsou. Ve svém výzkumu použily tři literární korpusy a následně zkoumaly, jak předchozí úprava korpusů metriku ovlivňuje. Korpusy upravily třemi způsoby – celé texty, rozdělení na 300 a 500tokenových segmentů a modelování pouze předem určených slovních druhů. Modely byly vytvořeny jak s neočištěnými, tak s očištěnými daty. Čištění sestávalo z odstranění slov s vysokou a nízkou frekvencí. Pro tematické modelování využily vědkyně algoritmu LDA z knihovny Mallet.

Po modelování vědkyně zjistily, že skóre koherence vykresluje typicky s vzrůstajícím počtem témat čtyři tendence: stoupající, klesající, parabolickou a nespojitou. Stoupající tendenci měla skóre koherence nesegmentovaných neočištěných dat. Klesající naopak nesegmentovaná očištěná data. Parabolickou křivku měla skóre koherence segmentovaných očištěných i neočištěných dat. Nespojitou tendenci pak měla data z jednoho určitého korpusu. Většina hodnot se nacházela v rozmezí mezi 0,43 a 0,5. Obecně lepší skóre koherence vycházelo s očištěnými texty a pouze s předem určenými slovními druhy.

Vědci pod vedením Jorise J. van Zunderta se zaměřili na hypotézu, zda témata

románů korelují s žánrem románu.[101] Svým výzkumem navázali na výzkum vědce

Christofa Schöcha.[102] Na rozdíl od ostatních zahraničních prací vědci van Zunderta pro tematické modelování použili algoritmus Top2Vec, který je novější než LDA nebo NMF. Jejich korpus sestával z více než 10 000 holandských románů vydaných mezi lety 2009 a 2019. Součástí korpusu byla i informace o žánrovém zařazení románů, díky které mohli tematické zařazení porovnat.

Před tematickým modelování vědci nejprve texty upravili tak, že z nich vyřadili vlastní jména a texty lemmatizovali. Následně vyřadili nejvíce a nejméně frekventovaná slova. Pro samotné modelování vyzkoušeli dva přístupy – rozdělení textů knih na 5000tokenová okna a použití celých textů knih. Následně pak otestovali, jak témata modelu Top2Vec korelují s žánry. Zjistili, že téma modelu Top2Vec silně koreluje s žánrovým zařazením, a to v obou přístupech. Dále pak zjistili, že některá témata se vytvořila kolem děl jednoho autora. Z toho usoudili, že se buď autor drží žánrového zařazení, nebo by bylo potřeba se více zabývat tím, co znamená téma v souvislosti s literární vědou.

V českém prostředí se tematickému modelování věnovaly dvě závěrečné práce z Českého vysokého učení technického. Ty využily již zmíněného projektu o motivických klastrech *Korpusu českého verše.* Jedná se o bakalářskou, resp. diplomovou práci Anny

Tesaříkové a Martina Bendíka.[103] Oba se ve své práci vytvořili tematické shluky z básní 19. a počátku 20. století. K tomu využili několik metod, které pak navzájem porovnali. Anna Tesaříková použila pouze tzv. metody učení bez učitele, tedy metody, které nepotřebují informace o klasifikaci dat. Mezi metody zařadila např. latentní sémantickou analýzu (LSA), BERT, Latentní Dirichletovu alokaci (LDA) či Top2Vec. Poslední dvě zmíněné metody využívám i já v této práci. Martin Bendík naproti tomu využil jak metody učení bez učitele, tak s učitelem. Kromě již zmíněných metod učení bez učitele použil klasifikátory naivní Bayes či SVM, které spadají do kategorie učení s učitelem a potřebují tak, aby alespoň část dat (zvaná jako trénovací data) byla anotovaná. Básně si autor rozdělil do 25 předem daných témat s tím, že rozlišil, zda je jednotlivé téma v básni

přítomno více, méně, či zda v básni není přítomno vůbec.[104] K anotaci si vybral 500

náhodně vybraných básní, což je přibližně 0,75 % básní z celého korpusu.[105]

Klíčové pro závěry obou autorů byly dvě metriky algoritmů – koherence a diverzita. Koherence měří příbuznost slov v daném tématu, diverzita pak to, jak moc jsou od sebe témata vzdálená. Autor i autorka se shodli, že nejlepší metodou je Top2Vec, který

vykazoval nejčitelnější výsledky.[106] Nejhorší výsledek podle Martina Bendíka vykazoval

algoritmus LDA,[107] podle Anny Tesaříkové modely BERT a LSA,[108] které Martin Bendík nepoužil. Metody učení s učitelem pak vydávali obecně horší výsledky než metody učení

bez učitele. Podle Bendíka to může být kvůli tomu, že procento anotovaných básní bylo

velmi malé oproti celému korpusu.[109]

V tomto kontextu je také potřeba zmínit diplomovou práci Julie Klimentové, která

vznikla na Ústavu informačních studií a knihovnictví FF UK.[110] Práce se zabývá topic modellingem frankofonních rapových textů. Svůj korpus Klimentová získala z veřejně dostupných zdrojů a celkově sestával z 5 061 francouzsky rapovaných písní. Klimentová k modelování využila algoritmus LDA, k výběru nejlepších parametrů použila kombinaci

skóre koherence a perplexity. Ve výsledku tak vybrala model, který měl 16 témat, skóre

koherence vyšlo na 0,311 a perplexita na -13,666.[111]

Loňský rok (2022) vyšla první publikace o digital humanities v českém prostředí: *Digitální obrat v českých humanitních a sociálních vědách,* která je souborem prací

zabývajících se digital humanities v českém kontextu.[112] Její součástí je i práce Radima

Hladíka o modelování témat v české sociologii.[113] Jako korpus použil *Sociologický časopis*, který je veřejně dostupný v repozitáři jazykových dat LINDAT/CLARIAH-CZ. Ten sestává z 522 článků od českých autorů a autorek. U 499 šly zpětně dohledat citační údaje. Hladík pro modelování použil algoritmus LDA v jazyce R, který na začátku potřebuje definovat počet témat. Parametr počtu témat byl na základě sekundární literatury nastaven

na 35.[114] Autor pak nalezená témata do hloubky analyzoval. Zajímavých zjištěním bylo rozložení podílu individuálního a kolektivního autorství. U témat předpokládajících důraz na empirické a kvantitativní přístupy tvořilo kolektivní autorství polovinu i více publikačního výkonu. Naopak historizující a esejistická témata tvořilo kolektivní autorství

méně než desetinu publikačního výkonu.[115] Hladík obdobně v tématech zanalyzoval i podíl

autorů a autorek podle genderu.[116]

Metoda tematického modelování přináší zajímavý vhled do literárních korpusů. Všechny zmíněné výzkumy nepředstavovaly výsledky tematického modelování samy o

sobě, ale ve srovnání s daty anotovanými od literárních vědkyň a vědců,[117] porovnávaly je

s dalšími ukazateli jako je gender,[118] zkoumaly metriku skóre koherence,[119] nebo

srovnávaly témata vzhledem k žánrovému zařazení výchozích textů.[120] Lze proto říci, že výsledky tematického modelování zatím nelze chápat jako rovnocenné závěrům klasické literární vědy, ale je stále potřeba zkoumat, jak se k nim komputační metody vztahují. Tematické modelování lze nicméně chápat jako metodu, která doplňuje závěry jiných metod literární vědy, což dokládá i já ve svém výzkumu.

5 Výzkumná metoda

V této kapitole se zabývám procesem přípravy dat a strukturou analyzovaného korpusu současné české prózy. Zaměřuji se na klíčové kroky, které byly nezbytné pro provedení tematického modelování. Prvním krokem bylo doplnění informací o věku a datu narození a úmrtí autorů, které nebyly obsaženy v původním korpusu. Následně jsem upravila samotný korpus, odstranila stop-slova a provedla rozpoznání pojmenovaných entit. Pro tento

výzkum použila veřejně dostupný seznam stop-slov od projektu CountWordsFree.[121] Seznam sestává z 463 slov. Vzhledem k velkému množství textů jsem segmentovala dokumenty na menší. Dále rozebírám proces vytváření tematických modelů Top2Vec a LDA, přičemž popisuji specifické kroky a volby provedené při implementaci obou algoritmů.

5.1 Personální autority

Součástí Českého národního korpusu není informace o věku či datu narození a úmrtí autora. Vhledem k tomu, že literární věda pracuje s generačním zařazením autorů a autorek, bylo potřeba tuto informaci ke korpusu přidat. Databázi národních autorit i s daty

narození a úmrtí poskytuje Národní knihovna ČR ve formátu MARC21.[122] Z této databáze bylo potřeba vyselektovat autory a autorky na základě jejich jména. Databáze však obsahuje velké množství jmen, která se opakují, proto bylo v případě konfliktních jmen vybráno vždy to nejpozdější datum narození. Některá data narození byla následně upravena, pokud bylo mechanicky přiřazeno nesmyslné datum narození.

5.2 Úprava korpusu

Český národní korpus svá data poskytuje ve formátu podobném CoNLL-U. Veškeré texty jsou tedy již tokenizované a lemmatizované, což usnadnilo práci s korpusem. Jak pro modelování pomocí algoritmu LDA, tak pro modelování algoritmem Top2Vec jsem použila lemmatizovaná slova. Součástí předzpracování textových dat bylo i vyřazení nejčastějších slov a slov, která obecně nenesou žádnou informaci. Obzvlášť algoritmus LDA, který nebere v potaz kontext a modeluje témata na základě statistiky, potřebuje z textu odstranit sémanticky nevýznamná slova.

Dále bylo potřeba z dokumentů vyřadit jména pomocí rozpoznávání pojmenovaných entit. Vlastní jména, jako např. jména postav, výsledky tematického

modelování zkreslují. K tomu jsem využila nástroj NameTag, který je přístupný přes

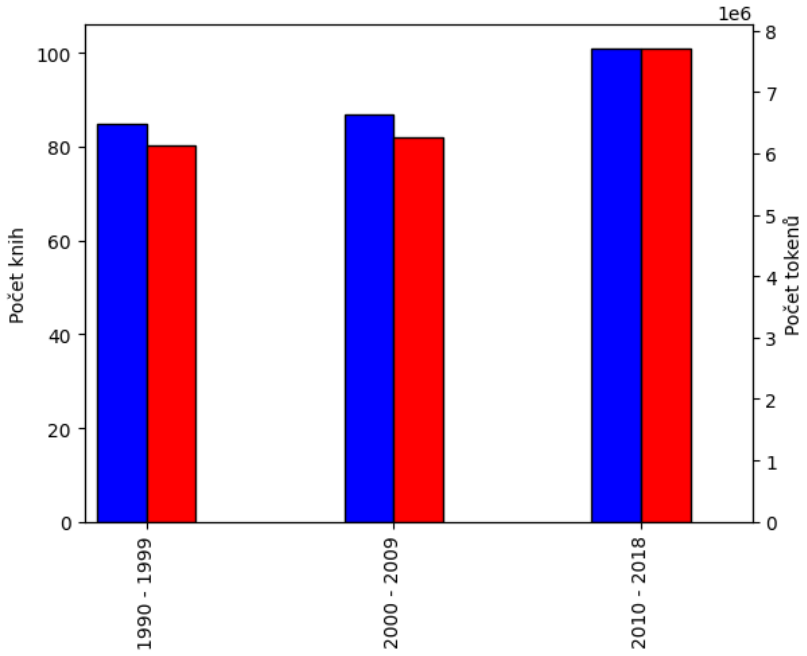
API.[123] Z dokumentů jsem vyřadila jména, která NameTag otagoval jako jména, „pf“, a

příjmení, „ps“.[124] Rozhodla jsem se jen pro jména a příjmení, jelikož na rozdíl od ostatních vlastních jmen spolu napříč knihami nesouvisí. Nástroj Nametag není stoprocentně spolehlivý, nicméně v hlavních tématech se téměř žádná vlastní jména neobjevila.

5.3 Skladba korpusu

Korpus pro analýzu jsem získala z veřejně přístupného korpusu SYN v9 současné psané češtiny, který kromě literárních děl obsahuje literaturu faktu a díla a články publicistického charakteru, které převažují. Díla a články pokrývají léta 1990 až 2018. Jelikož je korpus veřejně přístupný, musely být texty děl upraveny tak, aby z nich nebylo možné získat původní rukopis. Všechny texty proto byly rozděleny do bloků o maximální délce 100 slov. Tyto bloky byly následně promíchány, aby se z nich nedal původní text složit zpět.

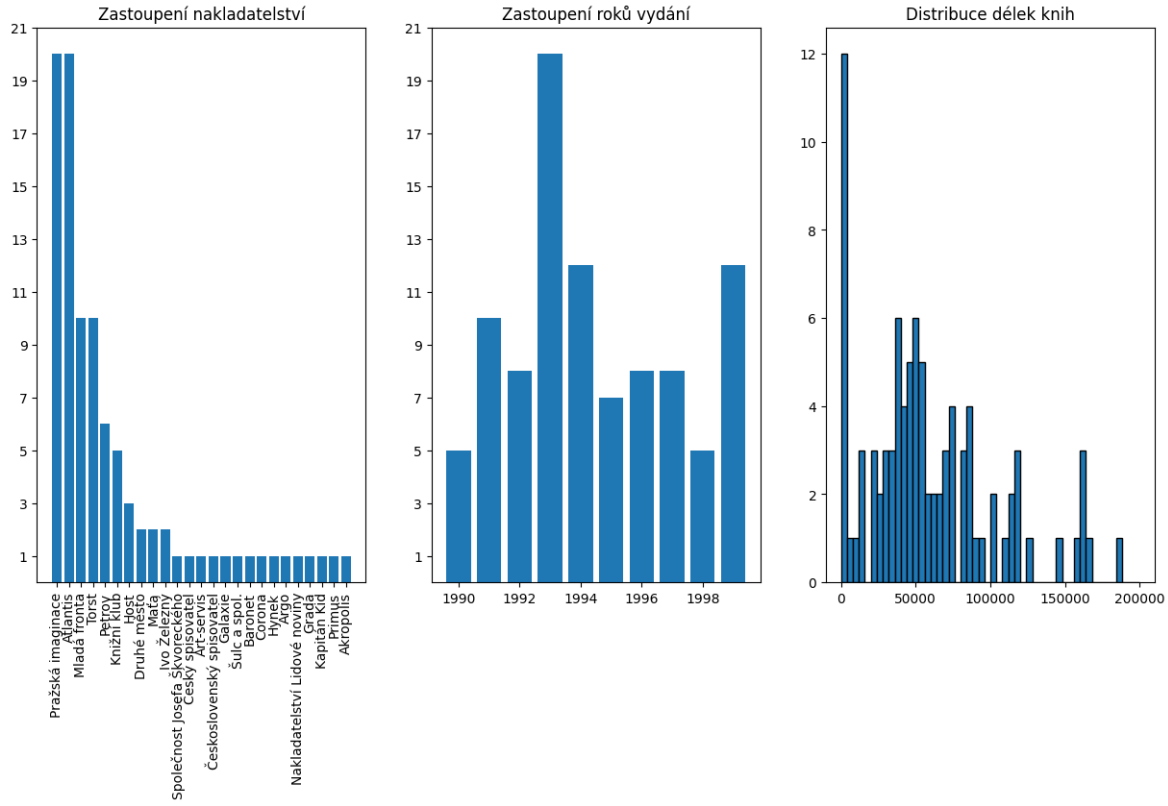
Pro účely literárního zkoumání jsem z korpusu vybrala knihy s označením „beletrie“ (FIC) a „próza“ – romány a novely (NOV). Tímto způsobem se do korpusu nedostaly žádné sbírky básní ani povídek, které jsou uloženy pod kódem „poezie“ (VER), resp. „kratší próza“ (COL). Ty by, vzhledem k tomu že texty jsou promíchány, narušovaly výsledky tematického modelování. Dohromady korpus sestává z 273 děl různého rozsahu. Dohromady mají všechny knihy 20 089 780 tokenů. Korpus byl rozdělen do tří subkorpusů podle data prvního vydání díla. První subkorpus obsahuje díla vydaná mezi lety 1990 až 1999, druhý sestává z děl vydaných mezi lety 2000 až 2009. Třetí subkorpus obsahuje díla, která poprvé vyšla mezi lety 2010 až 2018. Knihy vydané v roce 2019 Český národní korpus vůbec neobsahuje, proto je poslední zmíněný subkorpus složen pouze z knih vydaných mezi lety 2010 až 2018. Na následujícím grafu je vidět porovnání rozdělení počtu knih a počtu tokenů napříč rozdělenými subkorpusy. Modrý sloupec reprezentuje počet knih, červený sloupec pak počet tokenů v milionech. Z grafu je viditelné, že subkorpusy jsou přibližně stejně velké a obsahují přibližně stejný počet tokenů.



Obrázek 1: Grafické zobrazení počtu knih a celkového počtu tokenů

5.3.1 1990–1999

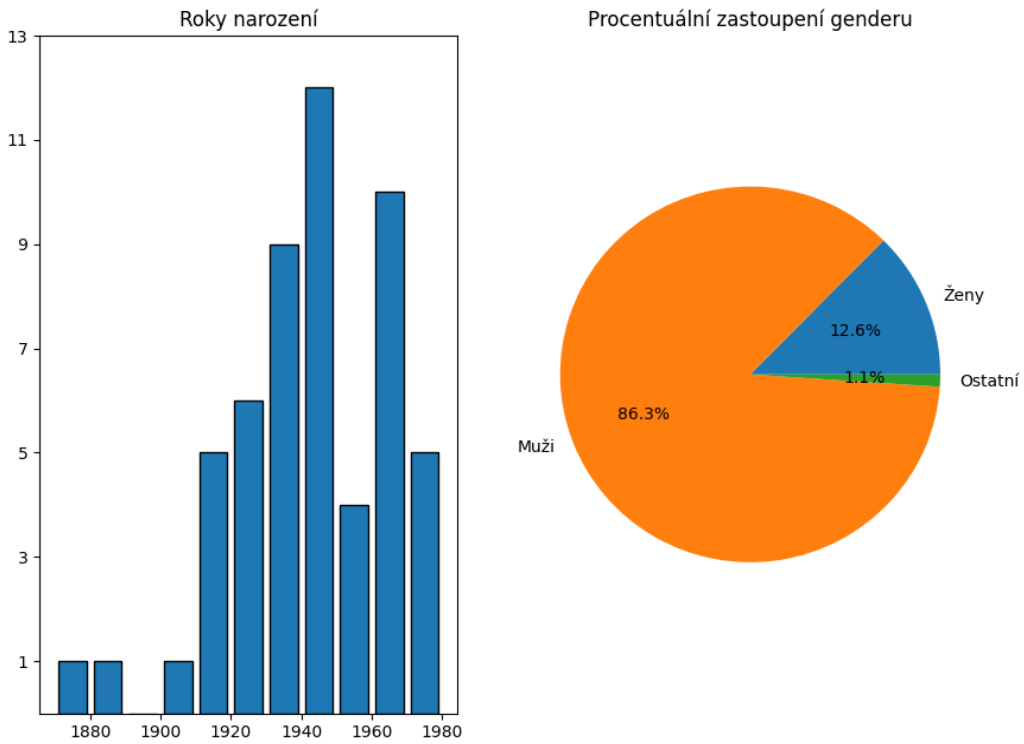
První subkorpus obsahuje 95 dokumentů. Některé tituly jsou však v korpusu rozděleny na několik částí. Dohromady tedy subkorpus obsahuje 85 titulů, které sestávají z 6 123 484 tokenů. Korpus ovšem bylo potřeba očistit o stop-slova a jména. Po očištění zůstalo v subkorpusu 2 472 310 tokenů. Knihy byly vydány dohromady ve 25 nakladatelstvích. Nejvíce jich bylo vydáno v nakladatelství Pražská imaginace a Atlantis, v každém vyšlo po dvaceti knihách. Z roků vydání je nejvíce zastoupen rok 1993, kdy bylo vydáno 20 knih. Nejméně pak v letech 1990 a 1998. Na následujících grafech je též vidět distribuce délek knih v subkorpusu děl z devadesátých let. Subkorpus obsahuje 12 děl, velmi krátkého rozsahu. Konkrétně jde o povídky Bohumila Hrabala, které vyšly jako samostatné svazky v Pražské imaginaci.



Obrázek 2: Statistiky děl subkorpusu 1990–1999

Na rozdíl od ostatních dvou subkorpusů je v subkorpus z let 1990 až 1999 menší rozmanitost autorů. Celkově je jich v subkorpusu 57, z čehož od několika nejznámějších autorů je subkorpusu vícero titulů. Nejvíce ze subkorpusu vystupuje Bohumil Hrabal, od kterého je v subkorpusu celkově 14 titulů. Několik těchto titulů jsou nicméně povídky vydané jako samostatné svazky. Subkorpus také obsahuje tři díla Ladislava Klímy, která poprvé vyšla v letech 1991, resp. 1996. Po zvážení byla díla v subkorpusu ponechána i přes to, že byl Ladislav Klíma v té době už přes 60 let po smrti. V devadesátých letech však vyšlo několik děl, která si autoři psali „do šuplíku“. Nejsou to tedy díla, která by vznikala v devadesátých letech.

Ze všech tří dekád je v 1. subkorpusu největší rozptyl roků narození autorů. Nejdříve narozený autor v subkorpusu je Ladislav Klíma, který se narodil v roce 1878. V devatenáctém století se pak narodil ještě jeden autor, který je v korpusu zastoupen – Jaroslav Durych, který se narodil v roce 1886. Naopak nejpozději se v subkorpusu narodil autor Pavel Hájek (1977) a Václav Bartuška (1975). Celkový podíl žen-autorek je v subkorpusu pouze necelých 13 %. Subkorpus také obsahuje dílo *Totální brainwasch* od dvojice autorů Josefa Vadného a Zdeničky Spruzené, což je pseudonym autora Tomáše Mazala. Ti jsou do statistiky zaneseni jako „ostatní“.

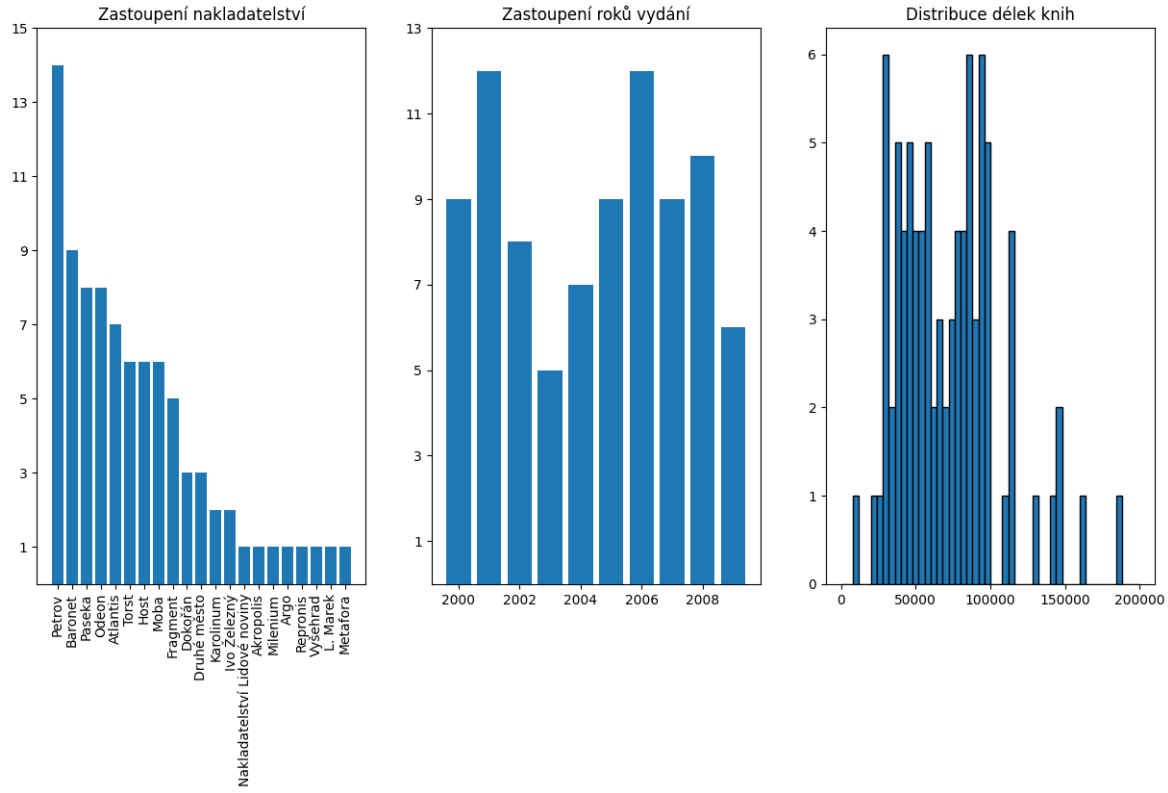


Obrázek 3: Statistiky autorů a autorek děl subkorpusu 1990–1999

5.3.2 2000–2009

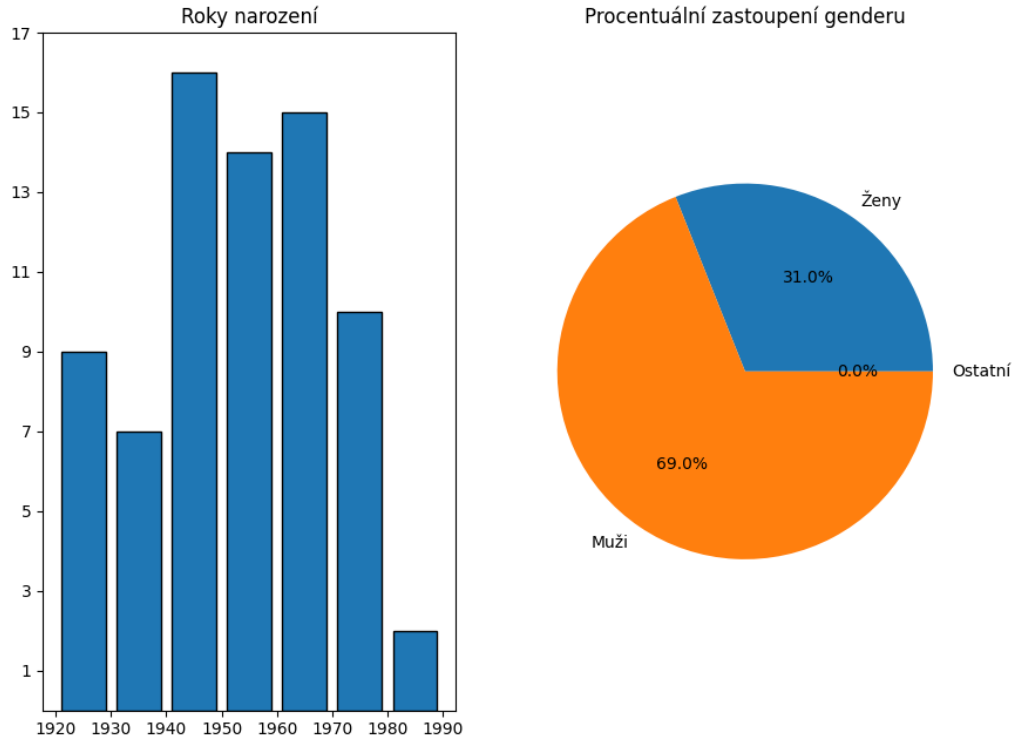
Druhý subkorpus obsahuje 87 titulů, které dohromady čítají 6 255 262 tokenů. Po očištění stop-slov a vlastních jmen má subkorpus 2 491 650 tokenů. Všechny tituly v subkorpusu jsou od celkově 73 autorů, což je o 16 autorů více než v předchozím subkorpusu. Pouze od autorky Moniky Zgustové jsou v subkorpusu 3 tituly, od ostatních autorů a autorek jsou v subkorpusu maximálně dva nebo (převážně) jeden titul. Knihy vyšly v celkově 21 nakladatelstvích. Nejvíce (celkově 14) jich vyšlo v nakladatelství Petrov (pak 9 titulů v nakladatelství Baronet). Nejvíce knih (po čtrnácti titulech) vyšlo v letech 2001 a 2006,

nejméně naopak v roce 2003 (pouze 5 knih). Oproti subkorpusu devadesátých let jsou v subkorpusu první dekády jednadvacátého století v průměru delší díla. Jejich průměrná délka se pohybuje mezi necelými 50 000 a 100 000 tokeny.



Obrázek 4: Statistiky děl subkorpusu 2000–2009

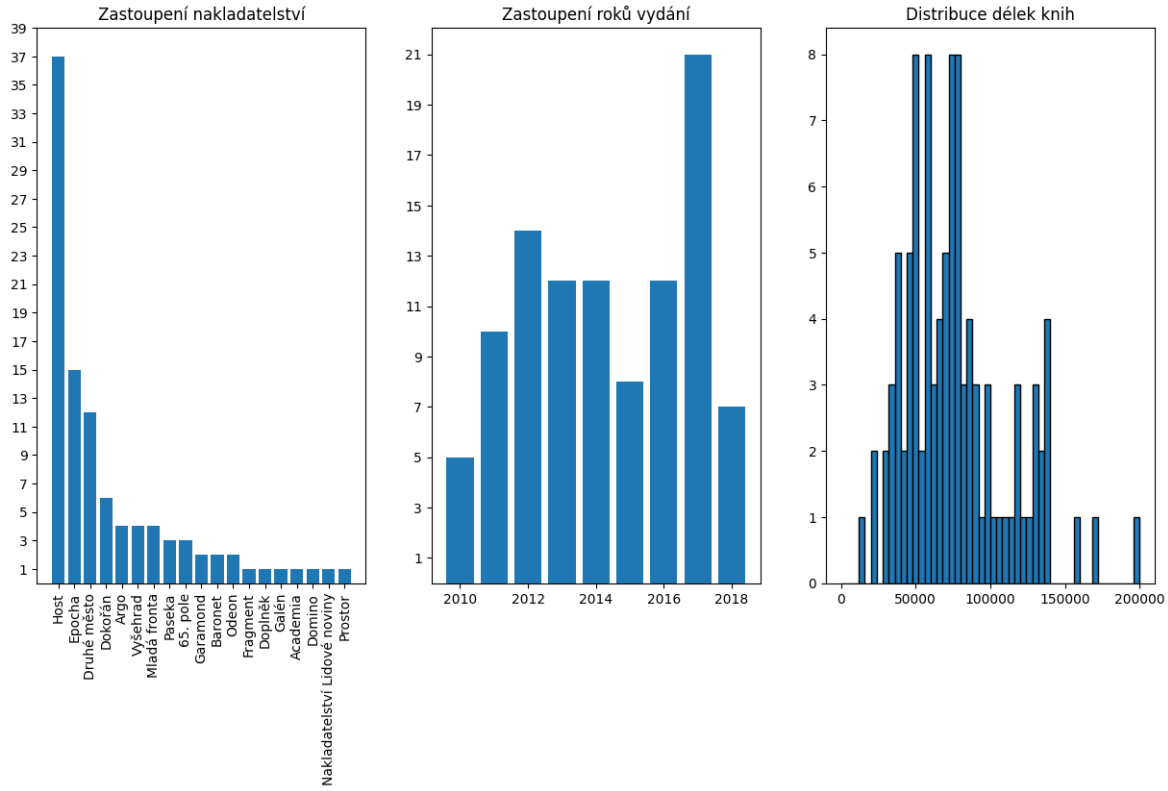
V subkorpusu první dekády jednadvacátého století je nejdříve narozeným autorem Zdeněk Rotrekl (1920). Po něm následuje autor Karel Eichler (1921). Mezi nejpozději narozené autory naopak patří autorka Petra Soukupová (1982) a autor Jan Jícha (1980). Z celkového počtu titulů je v subkorpusu 31 % napsáno autorkami. Subkorpus z let 2000 až 2009 se také vyznačuje tím, že má oproti ostatním subkorpusům nejmenší rozptyl co do nakladatelství a roků vydání.



Obrázek 5: Statistiky autorů a autorek děl subkorpusu 2000–2009

5.3.3 2010–2018

V posledním subkorpusu z let 2010 až 2018 je dohromady nejvíce knih – 101. Celkově čítají 7 711 034 tokenů, po očištění pak 3 085 973. Knihy byly napsány celkově 91 autory a autorkami. Nejvíce titulů (celkově 3) bylo napsáno autorkou Biancou Bellovou. Ostatní autoři a autorky jsou zastoupeni po jednom nebo dvou dílech. Nejvíce knih (21) vyšlo v roce 2017, nejméně (5) v roce 2010. Ze všech nakladatelství je nadreprezentováno nakladatelství Host, ze kterého je v subkorpusu 37 knih.



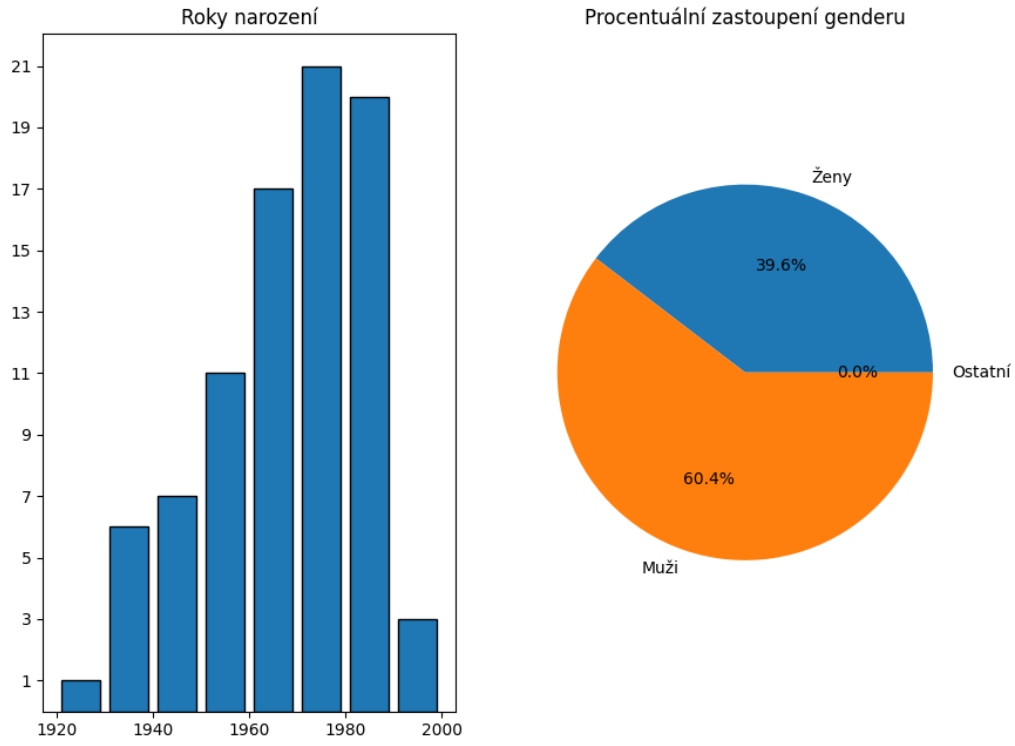
Obrázek 6: Statistiky děl subkorpusu 2010–2018

Podle literární vědkyně Aleny Šidákové Fialové se autorky vydávající v Hostu vyznačují „čtenářsky vstřícn[ým], dějově atraktivním příběh[em], […] silný[mi] ženský[mi] postav[ami] a témat[e]m partnerských a rodinných trápení a odhalování

(dávných) traumat“.[125] Z celkových 37 knih bylo 17 z nich napsáno ženami autorkami. To je téměř polovina z celého nakladatelství, tedy více, než je celkový průměr autorek v subkorpusu. Dá se tedy očekávat, že se „hostovská“ škola autorek promítne do závěrů.

Do korpusu se také dostaly dvě knihy, které svým zařazením sedí spíše do literatury faktu. Je tedy sporné, zda do korpusu knihy zařadit. Vzhledem k tomu, že biografie, žánr na pomezí literatury faktu a fikce byl (a stále je) v posledním desetiletí velmi populární, knihy byly nakonec v korpusu ponechány.

Třetí subkorpus obsahuje díla autorek a autorů, kteří se narodili mezi lety 1929 až 1991. Mezi nejdříve narozené autory v subkorpusu patří Miloš Hoznauer (1929) a Ota Filip (1930). Subkorpus také naopak obsahuje tři osoby, které se narodily až v devadesátých letech: Adam Skořepa (1991), Tereza Matoušková (1990) a František Kaleda (1990). Zdaleka nejvíce autorů a autorek se narodilo mezi šedesátými a osmdesátými lety. V subkorpusu druhé dekády jednadvacátého století je nejvíce děl z korpusu – přes 35 % – napsáno ženami.



Obrázek 7: Statistiky autorů a autorek děl subkorpusu 2010–2018

5.4 Modelování

Pro tematické modelování jsem zvolila dva popsané modely – LDA a Top2Vec, konkrétně jejich implementaci v knihovnách jazyka Python. Pro model LDA jsem použila

implementaci z balíčku gensim.[126] Implementace tohoto balíčku dovoluje uživateli krom počtu témat zvolit mnoho dalších parametrů. Dva nejzákladnější parametry algoritmu

LDA, „alpha“ a „beta“[127] jsem nastavila na „auto“, to znamená, že se algoritmus parametr naučí sám z korpusu. Dále jsem nastavila parametr „passes“ na 10. Tento parametr ovlivňuje, kolik částí korpusu půjde na trénování. Nastavení na 10 je standardním nastavením tohoto parametru. Pro modelování jsem využila pouze podstatná a přídavná jména, jelikož nesou nejvíce sémantických informací. Stejný balíček jsem využila i pro výpočet skóre koherence, který přímo implementuje výpočty z článku od vědeckého

kolektivu vedeného Röderem.[128] Implementace umožňuje vybrat, jakým způsobem skóre koherence spočítat. Já jsem pro výpočet použila výpočet *Cv*.

Implementace algoritmu Top2Vec je zatím dostupná pouze v jazyce Python v balíčku Top2Vec. Na rozdíl od LDA není potřeba modelu nastavovat žádné další parametry. Pro modelování jsem vyzkoušela jak korpus se stop-slovy, tak bez nich. Pro oba jsem vypočítala skóre koherence a následně porovnala. Modelu trénovaném na korpusu očištěném o stop-slova vycházelo v průměru vyšší skóre koherence, proto dále v práci pracuji pouze s korpusem očištěným od stop-slov.

Jelikož ani jeden algoritmus nezvládne zpracovat celé knihy najednou, dokumenty bylo potřeba rozdělit na několik stejně velkých segmentů. Pro každý model jsem vyzkoušela rozdělení na segmenty o 100, 500, 1000 a 2000 tokenů. Pro všechny jsem pak spočítala skóre koherence. Pro modely natrénované na segmentech o 100 a 500 tokenech nicméně vycházelo skóre koherence výrazně níž než pro ostatní dva typy segmentování. Do výsledků jsem proto nakonec započítala pouze modely natrénované na textech rozdělených na segmenty o 1000 a 2000 tokenech. Jak u modelu LDA, tak u modelu Top2Vec bylo přiřazení témat segmentům zprůměrováno napříč knihami. V případě modelu Top2Vec pak bylo knize přiřazeno téma, kterému byla kniha nejblíž.

Důležitým parametrem pro trénování obou modelů bylo nastavení počtu témat. Ten jsem nastavila na 10, 20 a 30 témat. Možností též bylo u modelu Top2Vec nechat počet témat na tom, které si model Top2Vec nalezne sám. Ten se ale, podle velikosti segmentů, pohyboval od poloviny počtu knih až po téměř celý počet knih v jednom subkorpusu. Znamenalo by to tedy, že by na každé dvě knihy existovalo jedno téma, což v souvislosti tohoto experimentu nebylo žádoucí.

Jelikož jsem též chtěla, aby výsledky subkorpusů byly navzájem porovnatelné, metriku skóre koherence jsem zprůměrovala napříč třemi subkorpusy. K tomuto kroku jsem se uchýlila i kvůli tomu, že počet knih je mezi subkorpusy velmi podobný. Znamená to však, že vybraný model nebyl vždy s nejvyšším skóre koherence. Veškeré kódy jsou k dispozici na mém githubu – https://github.com/panuscha/Czech-literature-topic-modelling.

6 Výsledky skóre koherence

V této části představím výsledky skóre koherence pro jednotlivé modely. Výsledky jsou prezentovány ve zkrácené podobě, úplné výsledky se nachází v příloze práce.

6.1 LDA

Algoritmem LDA bylo vytvořeno 6 modelů pro každý subkorpus – 3, kde byla díla rozdělena po 1000 tokenech a 3, kde se knihy rozdělily po 2000 tokenech. Počet témat byl u rozdělných trojic nastaven na 10, 20 a 30 témat. Ve výsledku tak pro každý subkorpus vznikly dva modely s 10 tématy, dva s 20 a dva s 30 tématy. Pro každý model bylo spočítáno skóre koherence. Pro jednoduchost byla skóre zprůměrována napříč subkorpusy. Výsledky skóre koherence byly zaneseny do následující tabulky. Jak můžeme vidět, metrika osciluje mezi 0,35 a 0,4. Výsledná témata nejlepšího modelu a skóre koherence pro všechny modely jsou v příloze.

Počet tokenů	**Počet témat**	**C_v**
1000	10	0,365
1000	**20**	**0,404**
1000	30	0,391
2000	10	0,358
2000	20	0,398
2000	30	0,391

Tabulka 1: Zprůměrované skóre koherence pro modely LDA

6.2 Top2Vec

Pro tematické modelování top2vec byly pro každý subkorpus namodelovány 2 modely – jeden, pro který byly jednotlivé knihy v korpusu rozděleny na 1000 tokenů a jeden, pro který byly rozděleny po 2000 tokenech. Každý model pak byl redukován na 10, 20 a 30 témat. Ve výsledku tak vyšlo 6 modelů pro každý subkorpus. Pro každý model bylo spočítáno skóre koherence. Pro jednoduchost byla tato skóre zprůměrována.

Na následující tabulce můžeme vidět, že v průměru vycházelo lepší skóre v tom případě, kdy byl model natrénován na větším počtu tokenů (tedy kdy celek obsahoval 2000

tokenů). Nejlépe vyšel model redukovaný na deset témat, který byl natrénován na celcích velkých max 2000 tokenů. Tento model jsem proto vybrala pro následnou interpretaci. Výsledná témata nejlepšího modelu a skóre koherence pro všechny modely jsou v příloze.

Počet tokenů	**Počet témat**	**C_v**
1000	10	0,447
1000	20	0,465
1000	30	0,471
2000	**10**	**0,479**
2000	20	0,470
2000	30	0,468

Tabulka 2:Zprůměrované skóre koherence pro modely Top2Vec

Jelikož Top2Vec modelu vycházela lepší skóre než modelu LDA, byl nejlepší model Top2Vec vybrán pro následnou interpretaci. Jak je vidno z tabulek výsledků, i Top2Vec model s nejhorším skóre koherence vycházel lépe než nejlepší model LDA. To není překvapivé zjištění, jelikož algoritmus Top2Vec pracuje s kontextem, kdežto algoritmus LDA pracuje se všemi slovy v textu najednou. K Top2Vec modelu byla navíc vytvořena projekce témat do 2D prostoru, která interpretaci subkorpusů doprovází.

Metriku skóre koherence použili i Anna Tesaříková a Martin Bendík, kteří

pracovali s Korpusem českého verše.[129] Jako první zkoumala tematické modelování Anna Tesaříková, které u modelu LDA vyšlo skóre koherence 0,43. Model Top2Vec vykazoval

skóre koherence 0,45.[130] Martinu Bendíkovi, který na práci Anny Tesaříkové navazoval, algoritmus LDA vyšel se skórem koherence 0,45, neredukovaný Top2Vec s 0,56 a

redukovaný dokonce 0,62.[131] Skóre koherence bylo jednou z rozhodujících metrik i pro pro Julii Klimentovou, která pracovala s korpusem frankofonního rapu. Pro tematické modelování použila model LDA. Nejlepší skóre koherence ji vyšlo 0,52. Metriku skóre koherence modelů LDA natrénovaných na literárních korpusech zkoumaly i vědkyně Inna

Uglanova a Evelyn Gius. Nejlepší výsledek, kterého jejich LDA model dosáhl bylo skóre

koherence 0,53.[132]

7 Analýza

V následující části popíšu výsledky obou algoritmů, zaměřím se však převážně na výsledky algoritmu Top2Vec, kterému vyšlo lepší skóre koherence. I proto popisuji výsledky modelu Top2Vec jako první a výsledky modelu LDA až jako druhé, které s prvními porovnávám. Do textu přikládám i kratší tabulku témat, delší přikládám do přílohy. Součástí témat Top2Vec je i mé pojmenování témat, které jsem vytvořila na základě slov v tématu, stručných anotací knih ve shluku či biografických textů o autorech a autorkách. Úplné výsledky jsou dostupné na mém githubu – https://github.com/panuscha/Czech-literature-topic-modelling.

7.1 1990–1999

7.1.1 Top2Vec

Skóre koherence pro první subkorpus vycházelo ve většině modelech nejlépe ze všech subkorpusů. Model, který byl vybrán pro interpretaci, měl skóre koherence téměř 0,52. Vysvětlení vysokého skóre se zřejmě nachází ve skladbě subkorpusu. Jak už jsem psala výše, subkorpus 90. let obsahuje více děl od stejných autorů. Tito autoři vytvořili samostatné shluky-témata, případně shluky, ve kterých jejich díla tvoří většinu tématu. Nejvíce titulů mají v subkorpusu autoři Michal Viewegh, Ludvík Vaculík a Bohumil Hrabal, kolem kterých se vytvořily témata 4, 5 a 8. Témata jsou proto pojmenována podle nich. Kolem děl autora se také vytvořilo téma č. 6 a 9, a to kolem Vlastimila Třešňáka a Ladislava Klímy. Většina segmentů knih také byla přiřazena ke stejnému tématu, případně většina segmentů k jednomu a jeden segment k jinému tématu. Výjimkou byla kniha Alexandry Berkové *Utrpení oddaného Všiváka*, jejíž segmenty byly přiřazeny ke čtyřem různým tématům.

Téma 1 (Normalizace, Vztahy)	**Téma 2 (Mafie)**	**Téma 3 (Krajina)**	**Téma 4 (Viewegh)**	**Téma 5 (Vaculík)**
Socialismus	jo	uvidet	usmev	decko
Dokazat	hele	zvedat	naprosto	chlapec
Pritel	tenhle	okno	roman	rukopis
Predevsim	prachy	ucitit	pohlednout	slunko

Zaroven	chlapek	okamzik	pochopitelne	takovyto

Téma 6 (Třešňák)	**Téma 7 (Historické Romány)**	**Téma 8 (Hrabal)**	**Téma 9 (Klíma)**	**Téma 10 (Autencitní literatura)**
times	turnovsky	cikanka	vzkriknout	celba
the	turnov	kracet	zvolat	teho
inu	grunt	lis	jmout	kurva
sajdkara	hrabe	balik	on	sichta
kriknout	vrchnost	vrchni	sestricka	mas

Tabulka 3: Prvních 5 slov témat modelu Top2Vec 1990–1999

Nejzajímavějším shlukem je bezpochyby shluk 8, který se vytvořil pouze z děl Bohumila Hrabala, jichž je v subkorpusu dohromady 14. Nejblíže jsou tématu knihy *Hlučná samota* a *Svatby v domě.* Obě mají kosinovou vzdálenost k tématu přes 0,47. Dále pak následuje kniha *Obsluhoval jsem anglického krále*, která je tématu vzdálena kosinovou vzdáleností 0,4. „Nejdále“ ke shluku je naopak povídka *Pohádka o zlaté Praze,* která je má od shluku vzdálenost 0,17. Shluk je tvořen slovy jako „cikánka“, „lis“, „balík“, „vrchní“ nebo „sklep“.

V tématu 4, který se zformoval ze všech děl Michala Viewegha v subkorpusu, nejvýrazněji vystupují *Účastníci zájezdu*, kteří mají kosinovou vzdálenost od téma téměř 0,39. To je patrné ze slov „zájezd“, „fotřík“ a „ségra“. Michal Viewegh je obecně považován za autora odpočinkové literatury, který tematizuje lásku, vztahy a každodenní život. Anna Fialová Šidáková i Lubomír Machala ho přiřazují do tzv. středního proudu literatury, tedy mezi autory, kteří se snaží nabídnout literaturu jak běžnému, tak erudovanému čtenáři. K tématu se připojili ještě další tři díla – *Ego* Ivana Matouška, *Jezdci pod slunečníkem* Romana Ludvy a *Vlak do Santa Fé* Jiřího Jiráně.

Téma 5 se zase koncentrovalo hlavně kolem Ludvíka Vaculíka. Na tématu se nejvíce projevila kniha *Jak se dělá chlapec,* což můžeme vidět na slovech „děcko“, „chlapec“ a „doma“. Jasně vystupuje i Ladislav Klíma, který kolem svého díla seskupil téma č. 9. Vzhledem k tomu, že díla autora pocházejí z počátku 20. století, je patrná použitá archaická jazyková forma, která se liší od ostatních témat.

Z témat dále vystupuje téma 2, které kvůli slovům jako „chlápek“, „kšeft“, „prachy“ a „policajt“ odkazuje na žánrovou literaturu. Většina slov je hovorová či slangová a odráží jazyk prostředí. Díla, která se nejvíce podílela na tvorbě tématu jsou *Sestra* a *Anděl* od Jáchyma Topola, *Vekslák 2* od Pavla Frýborta a *Dej mi ty prachy* od Ivy Pekárkové. Již z názvů je patrné, že knihy tematizují životní styl lidí, kteří se pohybují na hranici zákona. Díla jsou také často zasazena do 80. a obzvlášť 90. let. Slangové či až vulgární výrazy se také objevují v tématu 10. To však odráží převážně hornické prostředí dolů – „čelba“, „šachta“, „důl“, „předák“. Téma je nejvíce ovlivněno knihou *Pestré vrstvy* od Ivana Landmanna, autobiografickým románem z prostředí ostravských dolů a holandského exilu. I ostatní díla, ze kterých se téma vytvořilo, se dají přiřadit do skupiny děl která tematizují minulý režim.

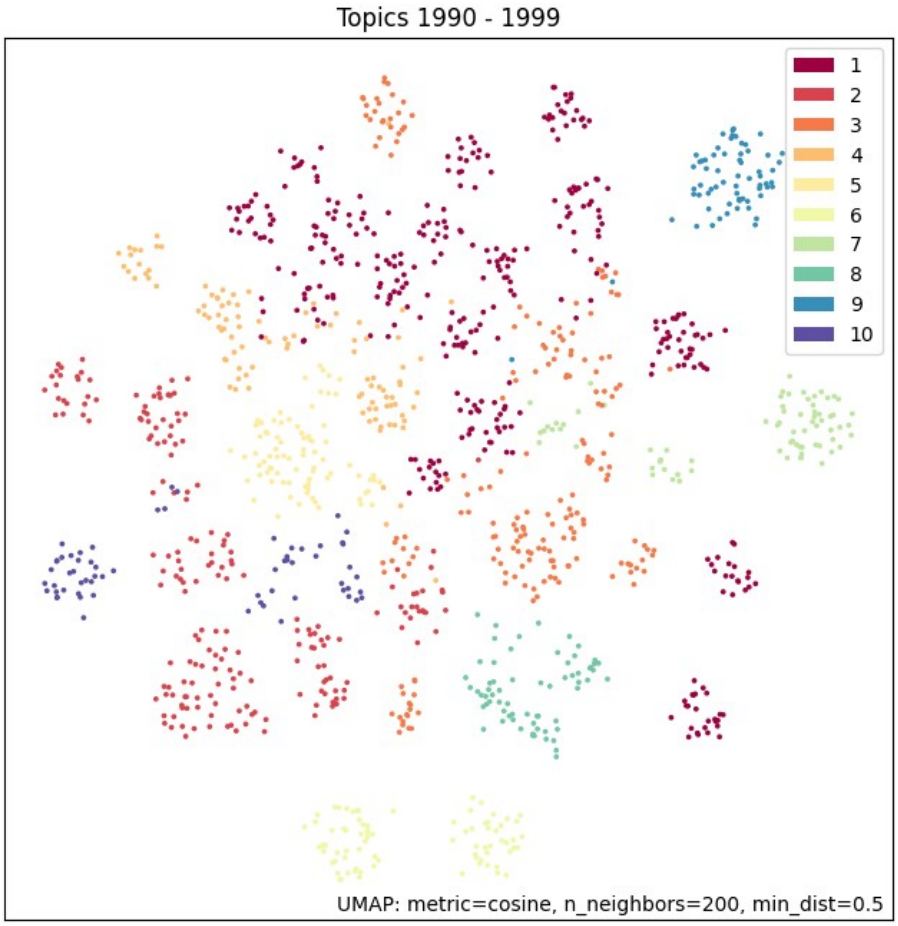
Zajímavé je i téma 7, které je jasně vymezeno historickou linkou subkorpusu. Je složeno z románů, které jsou zasazeny do historie a čerpají z dobových pramenů. Nejvýraznějším je první vydání rozsáhlého románu Jaroslava Durycha *Kouzelný kočár*, ve kterém autor mapuje příběhy svých předků.

Ženy autorky v některých tématech vůbec nejsou zastoupeny. Jelikož je subkorpus devadesátých let genderově nevyvážený a ženy tvoří jen 13 % korpusu, nelze usuzovat, že jsou některá téma predominantně „mužská“. Jsou to témata 3, 6, 7 a 8 a 9. U témat 6, 8 a 9 to není překvapivé, jelikož jsou témata vytvořena kolem jednoho autora. V tématu 7 jsou pouze 3 knihy a u tématu 3 předpokládám, že to, že autorky nejsou zastoupeny je zapříčiněno celkově nízkým výskytem žen v subkorpusu.

Kolem děl nejstaršího autora Ladislava Klímy narozeného roku 1878 se utvořilo téma 9, které je tvořeno pouze jeho díly. Následuje téma 8, které je tvořeno díly Bohumila Hrabala narozeného v roce 1914. Naopak téma 4 se vytvořilo kolem autorů, kteří se narodili nejpozději. Jejich průměrný rok narození vychází na rok 1960 s tím, že pouze jeden autor, Ivan Matoušek, se nenarodil v šedesátých letech. Dalším takovým tématem je téma 2, kde nejmladším autorem je autor Bohuslav Vaněk-Úvalský, narozený roku 1970. Naopak nejdříve narozeným autorem shluku je autor knihy *Gangsteři se nekoulují* Antonín Hodek, který se narodil v roce 1926.

Na následujícím obrázku 8 lze vidět zobrazení algoritmu dokumentů v 2D prostoru. Na první pohled je vidět, že všechny dokumenty tvoří jeden větší shluk, ze kterého vystupují dva menší shluky tématu 1, žluté téma 6, jeden větší shluk zeleného tématu 8 a modré téma 9. Žluté shluky tématu 6 jsou dvě knihy Vlastimila Třešňáka, zelený shluk

tématu 8 jsou všechna díla Bohumila Hrabala. Modrý shluk tématu 9 jsou tituly od Ladislava Klímy,



Obrázek 8: Zobrazení modelu Top2Vec pomocí UMAP 1990–1999

7.1.2 LDA

Jen málo témat algoritmu LDA vyšlo podobně jako témata algoritmu Top2Vec. Např. téma 16 u LDA se objevují slova jako „morče“, „inženýr“ a „chlapec“. Je to shluk, který se objevil i u algoritmu Top2Vec, téma 5, které se vytvořilo z knih Ludvíka Vaculíka. Obecně však témata spíše nekorespondují s tématy, které vygeneroval algoritmus Top2Vec. Oproti tématům Top2Vec se také témata zdají být lépe interpretovatelná a koherentní, navzdory nižšímu skóre korehence. Např. téma 4 slovy „ředitel“, „rodič“, „žák“ a „učitel“ jasně odkazuje na školní prostředí, téma 15 zase slovy „tábor“, „šaman“ nebo „náčelník“ příběhy z prostředí letního tábora pro děti.

Téma 1	Téma 2	Téma 3	Téma 4	Téma 5

Okno	doba	malý	manželka	město
Malý	duše	okno	ředitel	syn
Tělo	přítel	bůh	rodič	dcera
Noc	chlapec	Zeď	práce	doba
Stůl	malý	město	škola	babička

Téma 6	**Téma 7**	**Téma 9**	**Téma 8**	**Téma 10**
Doktor	peníze	práce	poručík	sovětský
maminka	doktor	město	město	agent
Pivo	stůl	Moskva	právo	tajný
Stůl	práce	dopis	císař	služba
Kluk	doba	Alexander	noc	zpráva

Téma 11	**Téma 12**	**Téma 13**	**Téma 14**	**Téma 15**
Kniha	hlavolam	český	přítel	šaman
komandant	puzzle	velvyslanec	srdce	stan
Papír	práce	grunžijský	římský	tábor
Balík	část	Praha	bůh	lovec
Lis	úsměv	rukopis	láska	náčelník

Téma 16	**Téma 17**	**Téma 18**	**Téma 19**	**Téma 20**
Morče	babička	specialista	autobus	komandant
inženýr	dědeček	autobus	žebřík	český
Stůl	učitelka	bůh	noc	jméno
Banka	divadlo	hospodář	ředitel	známý
Kolega	škola	Ježek	pouť	šavle

Tabulka 4: Prvních 5 slov témat modelu LDA 1990–1999

Jak už bylo popsáno výše, algoritmus LDA nepřiřazuje dílu pouze jedno téma, ale číselnou hodnotu vyjadřující, z kolika procent se z každého tématu dané dílo skládá. Výsledky tak nejdou jednoduše porovnávat. Lze se však podívat na díla, která mají k danému tématu nejblíže, tedy že procentuální zastoupení tématu je u nich majoritní. U

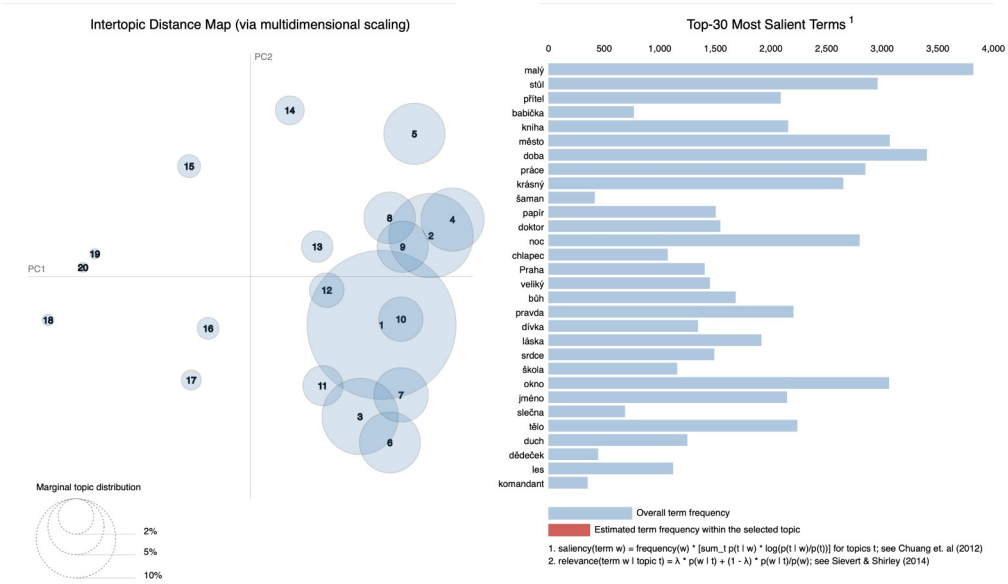
většiny témat existuje alespoň jedno dílo, ve kterém se dané téma objevuje z více než 90 %. Signifikantní je to u již zmíněného tématu 16, které tvoří knihu Ludvíka Vaculíka *Morčata* z 96 %. Všechny ostatní knihy jsou však tématem 8 tvořeny maximálně z 0,5 %. Znamená to tedy, že téma reprezentuje pouze knihu *Morčata*.

Díla Bohumila Hrabala byla modelem Top2Vec přiřazena do jednoho tématu, u modelu LDA tomu tak ovšem není. Téměř všechna autorova díla, až na *Love story,* jsou tvořena tématem 6, nicméně pouze knihy *Kůň truhláře Bárty, Modrý pondělí* a *Praha, město utajených infarktů* jsou tématem 6 tvořeny z více než 40 %. Je však třeba podotknout, že všechny tři knihy jsou tématem reprezentovány z více než 99 %. Pro ostatní jsou majoritní témata 11, 17, 19 a další.

Ani z děl Ladislava Klímy se nevytvořilo jasně definované téma jako u modelu LDA. Všechny autorovy texty jsou z nějaké části tvořené tématem 1, nicméně Klímovo dílo *Velký román* téma 1 reprezentuje pouze z 31 %. Tento text je z 58 % tvořen tématem 3. Podíváme-li se však na slova témat 3 a 1, zjistíme, že se některá slova („okno“, „malý“ a „město“) shodují. Témata jsou si tedy velmi blízká.

Na vizualizaci obrázku 9 je také vidět, jak velká témata jsou, jak daleko jsou od sebe a jak moc se překrývají. Téma 1 je viditelně největší, a není proto překvapením, že je přítomno u většiny knih. Dále vidíme, že témata 10 a 12 se nachází uvnitř tématu 1. S toho lze usuzovat, že se téma 10 a 12 přímo pojí s tématem 1. Tématem 10 je z 94 % tvořena kniha *Hřbitov vyzvědačů* Václava Pavla Borovičky, román o vyzvědačích KGB na území USA, Kanady a Británie. Tématem 12 je zase z 99 % tvořen román *Jezdci pod slunečníkem* Ludvy Romana. Jelikož jsou témata 10 a 12 v knihách tak silně přítomna, není možné, aby téma 1 bylo u nich též přítomno. U knih *Goldstein píše dceři* Ireny Douskové a *Květnové idy* Bohumila Hrabala, ve kterých je téma 10 přítomno z více než 50 %, je téma 1 rovněž přítomno, ale ne z více než 10 %.

Vizualizace též ukazuje menší témata 14 až 20, která jsou vzdálená od ostatních témat. Tématem 14 je převážně tvořena kniha *Ovidiova poslední láska* Jaromíry Sekotové, tématem 15 zase kniha *Šaman* Egona Bondyho. Tématy 18 a 20 není z většiny tvořena žádná kniha, jsou pouze přítomna v některých z děl.



Obrázek 9: Zobrazení modelu LDA 1990–1999

7.2 2000–2009

7.2.1 Top2Vec

Pro interpretaci byl vybrán model se skórem koherence téměř 0.46. Přesto jsou výsledky témat obecně srozumitelnější, než už subkorpusu 1990–1999. První jasně vystupující téma je téma 9 – druhá světová válka. Všechna slova se týkají Židů, koncentračních táborů nebo samotné války. Nejblíže k tomuto tématu má kniha *Zloděj kufrů* Arnošta Lustiga, která je zasazena do Terezínského tábora. Dále pak kniha Břetislava Olšera *... a Bůh osiřel*, která vypráví příběh chlapce, který zažil Osvětim a po válce se vydává do Izraele. Výrazné je také téma 6, které se vytvořilo z děl s historickou tématikou. Nejblíže jsou tomuto tématu kniha Vlastimila Vondrušky *Olomoucký bestiář*, která je zasazena do doby Přemysla Otakara II., a kniha *Zrození bestie* od Jany Švecové, která vypráví příběh Čachtické paní. Obě knihy měly kosinovou vzdálenost od tématu přes 0,39. Od autorky se do tématu přiřadila i její první kniha *Mé slzy zůstaly v Iráku*, která není zasazena do dávnější historie, nýbrž do doby první války v Perském zálivu. Od tématu má kosinovou vzdálenost přes 0,25. Domnívám se, že v tomto případě styl autorky převážil nad tématem knihy a obě její díla spojil dohromady.

Téma 1 (Rodinné	Téma 2 (Detektivní/	Téma 3 (Dětský příběh)	Téma 4 (Dějiny)	Téma 5 (Každodenní

vztahy)	**Krimi)**			**život)**
matcin	ponekud	trava	Zcela	jit
maminka	informace	koukat	Rochlice	pekny
laska	major	okolo	Tehdy	zase
okno	kontakt	spacak	Knihovna	jet
obraz	porucik	dlan	Znamy	vzit

Téma 6 (Historické romány)	**Téma 7 (Román pro ženy, oddychové čtení)**	**Téma 8 (Dospívání, vzpomínky na totalitní režim)**	**Téma 9 (2. světová válka)**	**Téma 10 (Zvíře a člověk)**
markyz	polibit	fakt	zidovsky	vlci
markyza	mobil	jo	arabsky	vlk
panos	pohlednout	hele	izraelsky	smecka
madame	mama	furt	jeruzalem	cosi
chlum	usmat	ponevadz	palestinec	dotknout

Tabulka 5: Prvních 5 slov témat modelu Top2Vec 2000–2009

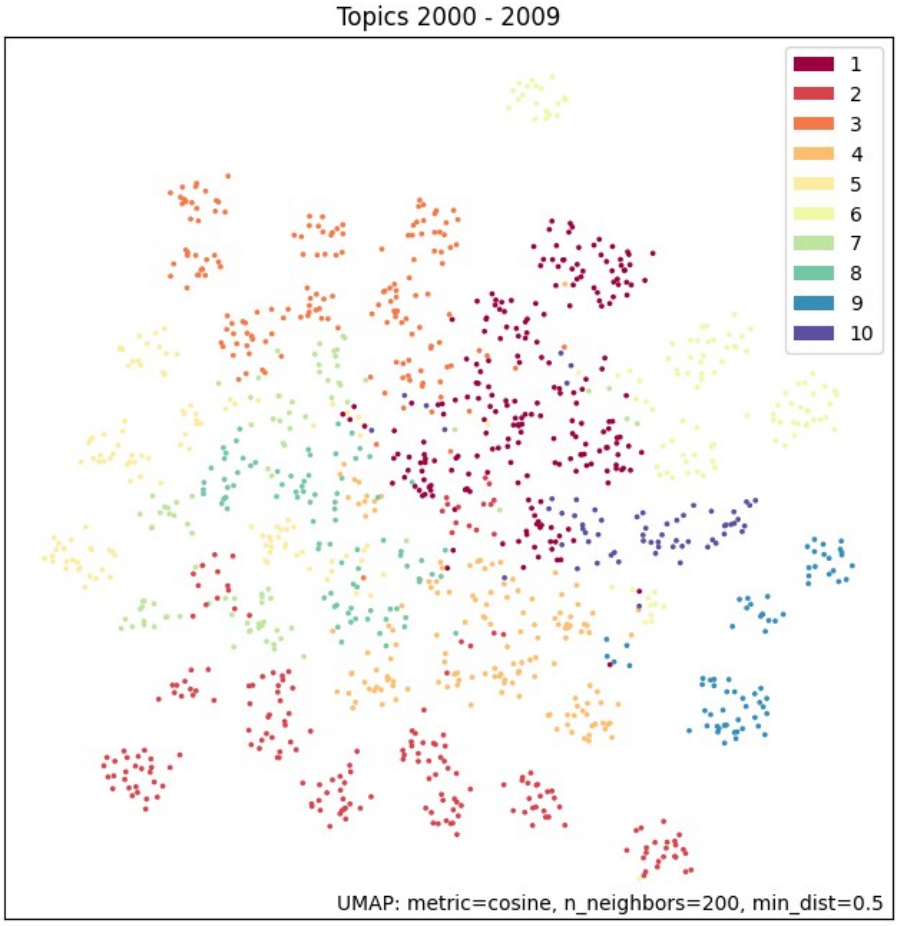
Téma 7 tvoří shluk kolem slov označujících intimní vztahy a emoce. Vystupují slova jsou např. „políbit“, „usmát“, „rozesmát“ či „zadívat“. Téma se nejvíce skládá z románů určených hlavně pro ženské publikum. Nejblíže se tématu blíží dílo autorky Romany Szalaiové *Tajná zahrada lásky,* která pojednává o vztahu televizní maskérky Aleny a jejího muže, architekta Robina. Od tématu má dílo kosinovou vzdálenost 0,42, téma tedy vystihuje celkem přesně. Tématu jsou blízká i dvě díla Michala Viewegha – *Román pro ženy* a *Případ nevěrné Kláry.* Obě díla se zabývají partnerskými a mileneckými vztahy.

Témata, která se v objevila jako první, jsou nejobecnější. Je proto obtížnější je interpretovat a nalézt mezi jejich komponenty spojující linku. To je patrné i z kosinové vzdálenosti jednotlivých děl k tématům. Např. k tématu č. 1 je nejblíže dílo *Zimní zahrada* Moniky Zgustové, která je od tématu vzdálena kosinovou vzdáleností necelých 0,29, což je méně než u již zmíněných témat. Většina děl z tématu 1 pracuje s širším dějinným kontextem a sleduje jednotlivé příběhy hrdinek a hrdinů, jejich milostné a rodinné vztahy atp., na jeho pozadí. To je patrné ze slov „matčin“, „maminka“, „láska“, „milovat“ a „otcův“. Celkově je tématu blízko 13 děl s průměrnou kosinovou vzdáleností 0,21.

Od žen-autorek je v tomto subkorpusu 31 % děl, což je výrazně více než v subkorpusu devadesátých let. V případě tohoto subkorpsu již můžeme říci, že autorky v některých tématech převažují a v některých zase zcela chybí. Téma č. 1 je převážně tvořeno díly autorek – 8 děl vs. 3 díla mužů-autorů. Ženy naopak zcela chybí v tématech 9 a 10.

Téma 9 se utvořilo kolem děl průměrně nejstarších autorů. Jejich průměrný rok narození je 1939. Druhé takové téma je téma 10 s průměrným rokem narození 1947. Naopak průměrný rok narození autorů děl z tématu 3 a 7 vyšel na rok 1966. V tématu 3 je pouze jeden autor z deseti autorů, Milan Charoust, který se narodil před rokem 1960. U tématu 7 se pak před rokem 1960 narodil pouze jeden autor z osmi, Martin Němec.

Stejně jako u subkorpusu devadesátých let je na zobrazení obrázku 10 v 2D vidět, že dokumenty vytváří jeden větší shluk a pár menších shluků, které stojí stranou. Jsou to hlavně texty s historickou tématikou, v nichž dominuje téma 6, a prózy Arnošta Lustiga, případně texty s tématikou 2. světové války obecně tvořící téma 9. Na okraji se též pohybuje shluk 2.



Obrázek 10: Zobrazení modelu Top2Vec pomocí UMAP 2000–2009

7.2.2 LDA

Skóre koherence modelu vyšlo na 0,39. Oproti subkorpusu devadesátých let zde u subkorpusu počátku milénia výsledky algoritmu LDA více odpovídají výsledkům algoritmu Top2Vec. Např. téma 8 modelu LDA je složeno z podobných slov jako téma 6 modelu Top2Vec. Jsou to slova jako „Chlum“, „královský“ nebo „panoš“. Tématem 8 modelu LDA je z více než 98 % tvořena kniha *Olomoucký bestiář* Vlastimila Vondrušky, ale překvapivě také kniha *Kloktat dehet* Jáchyma Topola. *Olomoucký bestiář* je tématu 6 modelu Top2Vec nejblíže, nicméně kniha *Kloktat dehet* byla Top2Vec modelem přiřazena k tématu 3. Téma 8 modelu LDA je také ze 79 % přítomno u knihy *Steiner* Martina Fahrnera, které model Top2Vec též nepřiřadil k tématu 6. Téma 8 modelu Top2Vec a téma 6 modelu LDA tak spojuje pouze kniha *Olomoucký bestiář*.

Téma 1	**Téma 2**	**Téma 3**	**Téma 4**	**Téma 5**
Malý	Malý	smrt	máma	Kluk
Doba	Okno	malý	doba	Doba
Okno	Stůl	bůh	práce	Soudruh
Noc	Doba	tělo	noc	Bratr
Stůl	Tělo	doba	malý	Máma

Téma 6	**Téma 7**	**Téma 8**	**Téma 9**	**Téma 10**
Malý	Kůň	Chlum	ocas	Dlaň
Auto	Pes	velitel	holka	Táta
Práce	Auto	tatínek	doba	Stůl
Noc	Les	královský	malý	Obličej
Holka	Kolo	panoš	slečna	Rameno

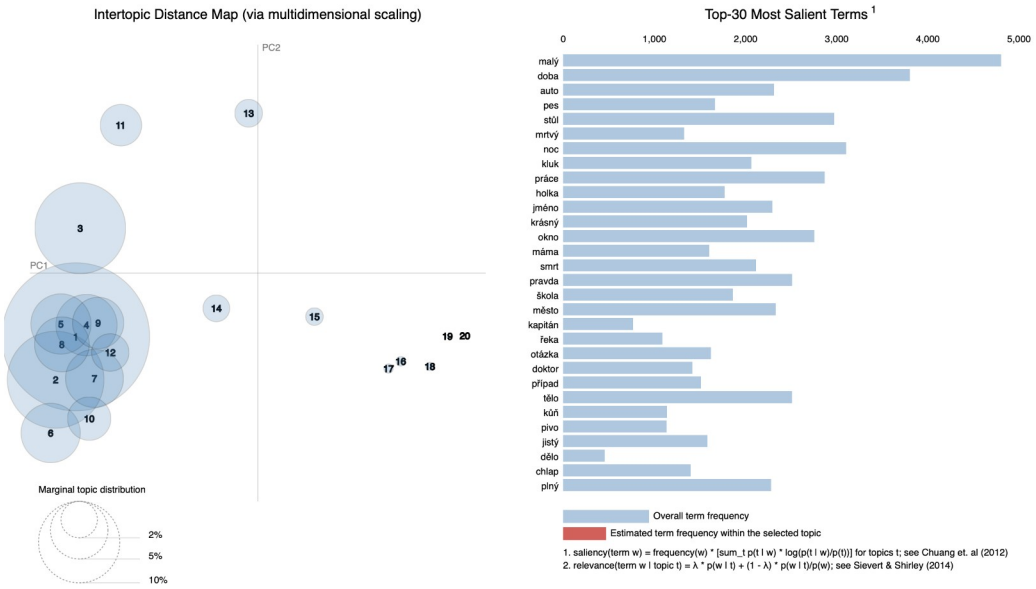
Téma 11	**Téma 12**	**Téma 13**	**Téma 14**	**Téma 15**
Dělo	Fišerová	družice	mrtvý	Dozorce
plukovník	Kluk	počítač	major	Smrt
kapitán	Kufr	kosmický	kapitán	Studna
Major	Holka	firma	pacient	Kniha
Sermonte	Noc	Brian	stůl	Luxemburgová

Téma 16	**Téma 17**	**Téma 18**	**Téma 19**	**Téma 20**
Doba	moře	pes	vzácný	Bezděk
Pravda	řeka	kolo	milostný	Síran
kamarád	malý	auto	planeta	Malý
Známý	pivo	přívěs	záchodový	Doba
Zub	chlap	krásný	malý	Noc

Tabulka 6: Prvních 5 slov témat modelu LDA 2000–2009

Dále pak téma 12 modelu LDA odkazuje na téma 9 modelu Top2Vec.[133] Tématem

Na vizualizaci obrázku 11 si lze všimnout, že většina témat si je velmi podobná a z tematických shluků vyčnívají pouze témata 11 a 13 až 20, která jsou ale mnohem menší než ostatní témata. U témat 11 a 14 je vždy jedno dílo, které je tématem tvořeno z více než 98 %. Další knihy tématy vůbec tvořeny nejsou. Témata 13 a 15 nejsou u žádných knih přítomna z více než 50 % a témata 16 až 20 se v knihách neobjevují téměř vůbec. Téma 20 je přítomno pouze u dvou děl – *Veselé příběhy podivné krásky* Karla Valáška a *Devětkrát jeden vrah* Anny Žídkové, a to maximálně z 0,25 %.



Obrázek 11: Zobrazení modelu LDA 2000–2009

7.3 2010–2018

7.3.1 Top2Vec

Skóre koherence zde vyšlo, stejně jako u subkorpusu 2000–2009, na 0,46. Ze všech témat jasně vystupuje téma 4 – komunismus. Téma je složeno ze slov jako „komunista“, „komunisticky“, „stranický“, „soudruh“, „socialistický“ atd. Z dalších slov tohoto tématu (maminka) lze také soudit, že je v daných dílech přítomna rodinná linka. Některé příběhy také mohou být vyprávěny očima dětského vypravěče nebo v nich je jinak přítomna rodinná tématika. Napovídají tomu slova jako „maminka“, „maminčin“, „škola“, „rodič“ a „tatínek“. Nejblíže je tématu 4 kniha *Volavčí sítě* autorské dvojice Luďka Navary a Miroslava Kasáčka, která pojednává o dvou studentech z Tišnova – Vlastimilovi Železném a Aloisi Pokorném, kteří byli zapojeni do třetího odboje. Od stejné dvojice k tématu přispěla i další kniha – *Příběhy třetího odboje*. V neposlední řadě je tématu 4

blízko i kniha *To je dost, žes zavolal* od Josefa Ejnara, která vypráví příběh chlapce od dětství jeho rodičů po první světové válce až do počátku sedmdesátých let minulého století.

Téma 1 (Tělesnost)	**Téma 2 (Historické romány)**	**Téma 3 (Rodinné vztahy)**	**Téma 4 (Komunismus)**	**Téma 5 (Prázdninové)**
Cítit	tma	jo	komunista	tata
pohnout	citadela	mama	maminka	strejda
pomalu	soumrak	gauc	komunisticky	ves
Ucítit	temnota	prachy	stranicky	chalupa
Upír	nebe	holka	soudruh	letos

Téma 6 (Mafie)	**Téma 7 (Psaní)**	**Téma 8 (Církev)**	**Téma 9 (Fantasy/ Středověk)**	**Téma 10 (Autorky Hosta)**
lobbista	spisovatel	kacir	Rytir	bohyne
Cching	roman	knez	panacek	kopanice
Sef	autor	bozi	podmesti	vlastovka
investor	literarni	cirkev	Mag	telo
Veta	napriklad	arcibiskup	brneni	policistuv

Tabulka 7: Prvních 5 slov témat modelu Top2Vec 2010–2018

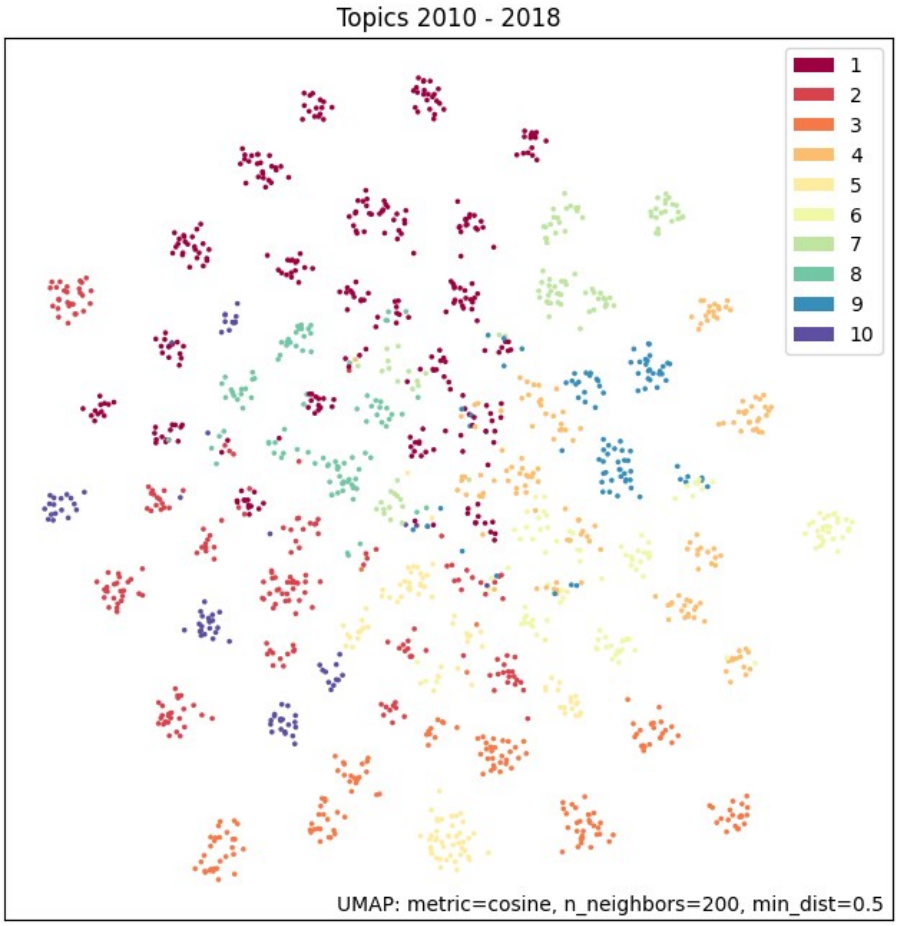
Téma 8 zase dalo dohromady historické romány. Slova nejvíce odkazují na církevní tématiku, což je patrné z prvních slov tématu „kacíř“, „kněz“, „boží“, „církev“ a „arcibiskup“. Nicméně i další slova v tématu odkazují na církevní tématiku. Nejblíže je tématu kniha *Svatý rváč* od Mileny Štráfeldové, historický román o osudech Jeronýma Pražského, církevního reformátora, který žil na přelomu 14. a 15. století. Kniha má od tématu 8 kosinovou vzdálenost přes 0,42. Znamená to tedy, že je kniha tématu velmi blízko. Druhá tématu nebližší kniha je próza Ivy Tajovské *Odpusť, že jsem se vrátil*, historický román zasazený do Československa po první světové válce, který vypráví příběh manželského páru – učitele, jenž se vrátil z italských legií, a jeho ženy. Dále je tématu 8 blízko kniha Oty Filipa *Valdštejn a Lukrecie,* historický román zasazený do přelomu 16. a 17.století, mapující osudy šlechtice Albrechta z Valdštejna.

V tématu 9 se zase shlukla slova týkající se rytířského středověku – „rytíř“, „brnění“, „přilba“, „zbraň“, zároveň se ale v tématu objevují slova odkazující k fantasknosti – „mág“, „magie“, „čarodějka“ a také boji – „kulka“, „zbraň“, „mrtvý“, „střílet“. Lze proto usuzovat, že téma je složeno z děl s historickou a fantasy tématikou.

Tématu 9 je nejblíže kniha Jaroslava Beznoska *Epos o panáčkovi*, akční sci-fi román o válce mezi „speciální zásahovou jednotkou rytířů“ a podsvětím. Kniha má od tématu 9 vzdálenost 0,4. Druhou nejbližší knihou s kosinovou vzdáleností 0,3 je fantasy román *Léky smutných* Pavla Gottharda o mladém studentovi, který se pomocí prášků na spaní dostává do fantazijního světa, kde rytíři bojují ve středověkých bitvách.

V subkorpusu druhé dekády jednadvacátého století je 40 děl od žen-autorek, což je přes 35 % subkorpusu. I v tomto případě lze říci, že se některá témata vytvořila převážně z děl autorek. Je to, stejně jako v subkorpusu první dekády, téma 1, kde ze 16 děl jsou jen 4 napsána muži. U tématu 10 je pouze jedno dílo ze čtyř napsáno mužem-autorem. Téma 7 je zase celé postaveno na dílech mužů-autorů. V tématu 5 je pak pouze jedno z 12 děl napsáno ženou, v tématu 9 pak pouze 2 díla z 9 napsáno ženami-autorkami.

Na rozdíl od předchozích dvou vyobrazení v 2D prostoru jsou na obrázku 12 zvýrazněné individuální knihy, které samostatně vytváří menší shluky. Na zobrazení tak není vidět jeden velký shluk. Z barev též nejsou příliš viditelná jednotlivá témata. Obarvené dokumenty v zobrazení ani nejsou u sebe, ale jsou vyobrazené po celém zobrazení. To je vidět např. u témat 9 a 10, u kterých se některé menší shluky objevují daleko od zbytku tématu.



Obrázek 12: Zobrazení modelu Top2Vec pomocí UMAP 2010–2018

7.3.2 LDA

Z prvního pozorování výsledků modelu LDA je patrné, že se některá slova opakují napříč mnoha tématy. Např. slovo „malý“ se objevuje v 11 tématech, což je více než polovina všech témat. Často se také opakují slova „okno“ – 6x a „tělo“ – 5x. Jednotlivá témata v subkorpusu se tak zdají méně odlišitelná.

Téma 1	**Téma 2**	**Téma 3**	**Téma 4**	**Téma 5**
tělo	malý	malý	Malý	táta
doba	Praha	tělo	Tělo	máma
kniha	škola	stůl	okno	malý
vlastní	doba	okno	město	auto
román	práce	rameno	ulice	holka

Téma 6	**Téma 7**	**Téma 9**	**Téma 8**	**Téma 10**
doba	bratr	základna	malý	rada
malý	loď	doba	práce	koule
sluneční	malý	skupina	Syn	ohnivý
šéf	král	Brno	auto	doba

rodina	tělo	archiv	doba	soudruh

Téma 11	**Téma 12**	**Téma 13**	**Téma 14**	**Téma 15**
táta	kůň	malý	sklo	chalupa
máma	malý	vlk	auto	stůl
doba	páter	kapitán	máma	ves
zeď	kolo	divoch	malý	okno
mrtvý	noc	vlastní	stůl	auto

Téma 16	**Téma 17**	**Téma 18**	**Téma 19**	**Téma 20**
student	plukovník	voják	řeka	holka
doba	ministr	rota	Kanaďan	řeka
stůl	Praha	mazák	Ptáčník	film
práce	auto	velitel	homunkulus	režie
kniha	vnitro	soudruh	doba	Paříž

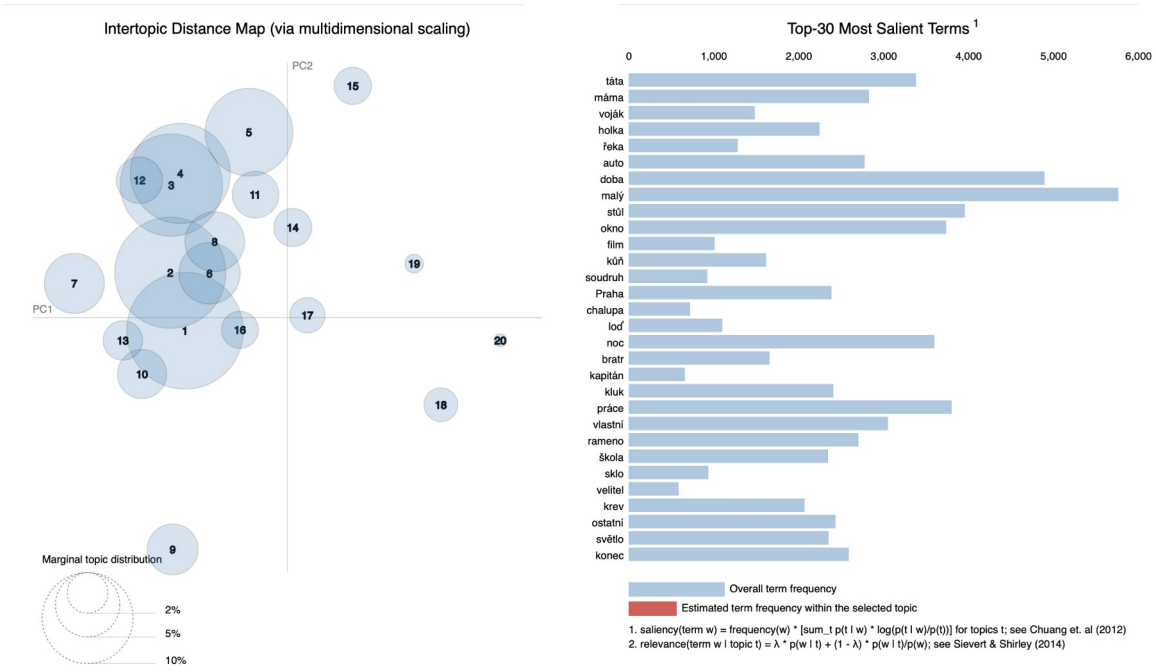
Tabulka 8: Prvních 5 slov témat modelu LDA 2010–2018

Stejně jako u subkorpusu první dekády i témata modelu LDA subkorpusu druhé dekády jednadvacátého století více odpovídají tématu algoritmu Top2Vec. Např. téma 15 modelu LDA částečně odpovídá tématu 5 modelu Top2Vec. V obou tématech se objevují slova jako „chalupa“, „ves“, „táta“ nebo „strejda“. Tématu 5 modelu Top2Vec jsou nejblíže knihy *Rybí krev* Jiřího Hájíčka a *Všechno je jen dvakrát* Michala Přibáně. Tématem 15 modelu LDA je zase z 74 % tvořena druhá Hájíčkova kniha v korpuse, *Dešťová hůl*. Rybí krev je tématem 15 tvořena z 48 %, z 50 % je nicméně tvořena tématem 5, kterým je z 93 % tvořena kniha Michala Přibáně.

Tématu 4 modelu Top2Vec částečně odpovídají témata 8 a 10 modelu LDA. Obě témata obsahují slova jako „soudruh“, téma 8 pak ještě „voják“. Knihy autorské dvojice Luďka Navary a Miroslava Kasáčka, které jsou tématu 4 nejblíže, jsou však podle modelu LDA tvořeny tématem 9, které obsahuje slova jako „základna“, „doba“, „tajný“ nebo „vězení“. Jsou to slova, která sice také souvisí s komunistickým režimem, nicméně žádné z nich samo o sobě ke komunistickému režimu přímo neodkazuje. Tématem 8 je nejvíce (z 90 %) tvořena kniha *Malá noční žranice* Jiřího Šimáčka, tématem 10 zase kniha *Tajemství ohnivých koulí* Miloslava Švandrlíka (z 96 %).

Na následující vizualizaci jsou témata vyobrazena v prostoru. Oproti předchozím vizualizacím modelů LDA jsou na této více definovaná témata. Prvních 5 témat je přibližně stejně velkých a také zabírají větší plochu. Vzdálenější témata od většiny ostatních jsou témata 9, 15 a 18. Jejich zobrazení má však větší průměr, než mělo

zobrazení minoritních témat u předchozích subkorpusů. Je zde také viditelný jasný překryv témat 3 a 4. Obě ve svých prvních čtyřech příčkách mají slova „malý“, „tělo“ a „okno“.



Obrázek 13: Zobrazení modelu LDA 2010–2018

8 Srovnání se sekundární literaturou

Ve druhé kapitole jsem zběžně shrnula tendence české polistopadové prózy, jak ji sepsali autorky a autoři souhrnů *V souřadnicích volnosti*, *V souřadnicích mnohosti* a dalších kratších statí. Současná česky psaná próza je zde rozdělena do několika proudů, které se vyznačují nejen tématem, ale i generačním či genderovým zařazením autorstva. Paralelu v tématy modelu Top2Vec jsem tedy nacházela na základě vícero znaků. V této části je též nutno zmínit, že v rámci této práce neprovádím žádné testy, pouze deskriptivně popisuji výsledky.

Díla, ze kterých *V souřadnicích volnosti*, *V souřadnicích mnohosti* a další kratší statě vycházely, se částečně liší od těch, která obsahuje *Český národní korpus*. Tento nedostatek otevírá interpretaci témat strojově odvozených z textů, které v sekundární literatuře nefigurují. Korpus *Českého národního korpusu* totiž nevybírá knihy na základě

kánonu, proto některá témata částečně odhalují proudy, které se zatím nepodílejí na tvorbě

akademické literární historie.[134]

8.1.1 Devadesátá léta dvacátého století

Lubomír Machala je toho názoru, že „[k]nižně publikované autobiografické deníky, vzpomínky či memoáry vytvářely v poslední dekádě dvacátého století jeden

z nejvýraznějších prozaických trendů“.[135] V textu zmiňuje Bohumila Hrabala a Ludvíka Vaculíka, kteří v rámci modelování vytvořili samostatné shluky 5 a 8. Oba autoři píšou o různých tématech, proto není překvapivé, že jsou v rozdílných shlucích. To, zda se oba inspirovali vlastním životem nelze algoritmicky zachytit.

Podle Lubomíra Machaly však autobiografické nebyly jen deníky, vzpomínky či memoáry, ale objevovaly se i útvary na pomezí autobiografie a prózy. Mezi taková díla řadí texty autorů Vlastimila Třešňáka nebo Ivana Landsmanna. Dva tituly Třešňáka – *Klíč je pod rohožkou* a *Evangelium a ostružina* – jsou součástí korpusu. Kolem dvou Třešňákových děl se vytvořil samostatný tematický shluk 6, z čehož lze usuzovat, že je styl Vlastimila Třešňáka relativně unikátní. Dílo Landsmanna – *Pestré vrstvy* se zase nachází ve shluku 10. Algoritmus Top2Vec k němu dále přiřadil díla *Hrdý Budžes* Ireny Douskové nebo *Totální brainwash* autorské dvojice Zdeničky Spruzené a Josefa Vadného. Slova v tématu 10 odkazují na autencitní literaturu.

Machala též tvrdí, že díky pracem Milana Kundery a Jana Křesadla došlo po

listopadu 1989 v českém prostředí ke zdomácnění pojmu postmoderna.[136] Postmoderna se však nevyznačuje tématem, nýbrž dalšími literárními prvky jako jsou hra se čtenářem,

specifický vypravěč „nezřídka […] vybavený[m] autobiografickými rysy“[137], metatextovost, intertextualita atd. Nelze tedy předpokládat, že algoritmus dokáže nalézt společné téma pouze na základě toho, že jsou díla chápána jako postmoderní na základě zmíněných formálních postupů. Díky tomu, že model Top2Vec pracuje nejen s frekvencí slov, ale i s kontextem, není nicméně vyloučeno, že model díla přiřadí ke stejnému tématu, shluku. Ze všech děl dvojice Kundera a Křesadlo je v korpusu pouze *Nesmrtelnost* Milana Kundery, která byla přiřazena k tématu 1.

K postmoderním autorům „střední“ generace, narozeným ve čtyřicátých a padesátých letech, pak literární historici řadí například Jiřího Kratochvila, Danielu

Hodrovou, Ivana Matouška nebo Michala Ajvaze.[138] Kromě Daniely Hodrové jsou všichni autoři zastoupeni v korpusu. Knihy Jiřího Kratochvila – *Uprostřed nocí zpěv*, *Avion* a *Noční tango* – byly algoritmem zařazeny k tématu 3, stejně jako dílo *Druhé město* Michala Ajvaze. *Ego* Ivana Matouška bylo naopak přiřazeno k tématu 4 k dílům Michala Viewegha. Průměrný rok narození autorů z tématu 3 vychází na rok 1950, což by odpovídalo autorům „střední" generace. 5 autorů, kteří napsali knihy nejblíže k tomuto tématu, se však narodilo buď před čtyřicátými lety, nebo až po letech padesátých. Nelze tedy říci, že by téma reprezentovalo literaturu postmoderních autorů střední generace. V tématu 4 je zase Ivan Matoušek jediný autor, který se narodil v rozmezí čtyřicátých a padesátých let.

Mezi poslední, nejmladší generaci postmodernistů řadí Machala autory Martina Komárka, Petra Ulrycha a Jáchyma Topola. Jsou to autoři, kteří se narodili v šedesátých letech. Jejich texty spojovaly s díly starších prozaiků kromě postmoderních východisek i

některá témata.[139] V korpusu jsou zastoupena pouze díla Jáchyma Topola z daného období – *Sestra* a *Anděl*. Obě díla byla přiřazena k tématu 2, které svými slovy a složením knih odkazuje spíš ke krimi žánru. Postmoderní postupy se nicméně vyznačují mícháním vysokých a nízkých literárních žánrů.

Dalším literárním proudem 90. let byl tzv. střední proud, který tvořil díla na pomezí náročné a populární četby. Do tohoto proudu řadí Lubomír Machala hlavně díla Michala Viewegha, Petra Šabacha, Romana Ludvu a Jana Jandourka. V korpusu jsou zastoupeny hlavně díla Michala Viewegha a dílo *Jezdci pod slunečníkem* Romana Ludvy, které bylo s ostatními díly Michala Viewegha přiřazeno do tématu 4. K autorům, kteří se snažili „o

propojení náročnější a čtenářsky přitažlivé literatury“[140] Machala navíc řadí díla Ivana Klímy a Pavla Kohouta. V korpusu je pouze dílo *Ten žena a ta muž* Pavla Kohouta, které bylo přiřazeno k tématu 1.

Lubomír Machala též zmiňuje, že i přes to, že „se opakovaně objevovaly názory, že

[…] historická próza [procházela] útlumovou fází, nelze tvrdit“[141], že by historická díla nevznikala. Díla, která Machala zmiňuje však nejsou zasazená do stejného období, proto není možné předpokládat, že by je algoritmus přiřadil do stejného tématu. Krom toho je ze zmíněných děl v korpusu přítomna pouze kniha *Poslední tečka za Rukopisy* Miloše Urbana, která byla přiřazena k tématu 7. Slova tématu jako „hrabě“, „vrchnost, „rod“ atd. opravdu odkazují k historickým románům. Jediný tematický shluk, který zatím nebyl zmíněn je shluk 9, který se vytvořil děl Ladislava Klímy, které prvně vyšly až v 90. letech.

8.1.2 První dekáda jednadvacátého století

Jak píše Alena Šidáková Fialová, „[v]ýrazný proud tvořilo v české próze nového milénia

[…] téma druhé světové války, holocaustu a života v protektorátu“.[142] Toto téma je jasně viditelné u shluku 9 v subkorpusu 2000–2010. Když se ale podíváme, jaké knihy ve svém textu Alena Šidáková Fialová cituje, ani jedna z nich se v tématu neobjevuje. Většina knih v korpusu vůbec není a dvě knihy jsou přiřazeny k jiným tématům – *Obyčejné životy* od Josefa Škvoreckého a *Jizvy* od Evity Naušové. *Obyčejné životy* byly přiřazeny k tématu 8, *Jizvy* zase k tématu 7. Alena Šidáková Fialová v kontextu děl zasazených do období druhé světové války zmiňuje dvě díla Arnošta Lustiga – *Dívka s jizvou* a *Případ Marie Navarové*, ani jedno sice není součástí korpusu, ten však obsahuje dvě jiné knihy od Arnošta Lustiga – *Zloděj kufrů* a *Nemáme na vybranou.* Obě knihy byly algoritmem přiřazeny do tématu 9.

Dále v první dekádě dvacátého století byly stále oblíbené postmoderní postupy,

které zdůrazňovaly vypravěčovu personu.[143] Alena Šidáková Fialová jako příklad zmiňuje autora Jiřího Kratochvila, od kterého je v korpusu kniha *Truchlivý bůh*. Ta byla algoritmem přiřazena k tématu 10. Dále autorka zmiňuje díla Jáchyma Topola – *Noční práce* a *Kloktat dehet,* které jsou součástí českého národního korpusu. Obě knihy jsou zasazeny od prostředí Československa po roce 1968. Algoritmus je přiřadil k tématu 3, který (podle anotací knih, které jsou tématu nejblíže) shlukuje knihy s dětským vypravěčem. Do stejného shluku bylo přiřazena i dílo autora Václava Chocholy *Děti z krechtů*, které tematizuje dětství za pozdní normalizace. O tomto díle Alena Šidáková Fialová píše v souvislosti s díly, která jsou zasazena do období sedmdesátých a osmdesátých let. Podle autorky tyto knihy tvořily „jeden z nejvýraznějších proudů nové

produkce“.[144] Knihy tematizující dospívání v normalizaci se shlukly i kolem tématu 8 – kniha Ireny Douskové *Oněgin byl Rusák* nebo knihy *Opilé Banány* a *Občanský průkaz* Petra Šabacha, které Alena Šidáková Fialová v kontextu normalizace zmiňuje.

Když se však podíváme na slova témat 3 a 8, nepoznáme, že se v tématu shlukují díla se zmíněnou tématikou. Ze slov jako „fakt“, „jo“, „hele“, „furt“, „táta“, „škola“ nebo „třídní“ sice můžeme odvodit, že knihy z tématu 8 jsou napsané hovorovým jazykem a tematizují školní prostředí, o normalizaci bychom se ale nedozvěděli, pokud bychom nevěděli, o čem knihy jsou. U tématu 3 je to ještě zřetelnější. Ze slov „okolo“, „spacák“, „chechtat“ nebo „kluk“ bychom nepoznali, že téma shlukuje knihy s dětským vypravěčem. Naopak téma 6 (historické romány) nebo 9 (2. světová válka) jsou ze shluků slov jasně rozpoznatelné.

Podle Aleny Šidákové Fialové bylo „[ú]středním tématem velké částí prozaické produkce […] mezilidské, zvláště milostné vztahy, čistě intimní sféra intimního života,

hledání svého místa na světě a mezi druhými“. [145] Dále zmiňuje, že se jedná o autorky mladší a nejmladší generace. Kolem rodinných a intimních vztahů by se ze slov témat zdálo, že se vytvořilo téma 1, nicméně když se podíváme na díla a autorky, které Šidáková Fialová zmiňuje, zjistíme, že k tématu 1 nebylo přiřazeno ani jedno z nich. V kapitole *Prózy soukromých dramat* Šidáková Fialová zmiňuje díla autorek Petry Hůlové, Petry Soukupové, Natálie Kocábové a Terezy Boučkové. Knihy *Rok kohouta* Terezy Boučkové a *K moři* Petry Soukupové se společně nachází v tématu 5. Ze slov tématu „chodit“, „jet“, „spát“, „jíst“, „jezdit“ lze usuzovat, že se v knihách odehrávají „každodenní“ činnosti. Většina autorů a autorek knih z tématu 5 (včetně Terezy Boučkové) však nepatří k mladší a nemladší generaci. Kniha *Cirkus Les Mémories* Petry Hůlové byla přiřazena k tématu 3.

Kromě próz, které líčily soukromé problémy, se podle Šidákové Fialové objevovaly

také dobově politické romány.[146] Z autorů a děl, které jsou v korpusu, zmiňuje badatelka knihy Ludvíka Vaculíka *Hodiny klavíru* a *Cesta na Praděd* a knihu Milana Exnera *Svatoušek.* Vaculíkova díla byla stejně jako *Rok kohouta* a *K moři* přiřazena k tématu 5, dílo Milana Exnera naopak k tématu 1. Dalším výrazným proudem se podle Šidákové Fialové staly příběhy zasazené do exotického prostředí. Ze zmíněných děl jsou však v korpusu přítomny pouze Hůlové *Cirkus Les Mémories* a *Zvlčení* Antonína Bajaji, který pojednává o nespoutaných vlcích žijících v karpatské divočině. Kniha je zařazena do tématu 10, ke kterému je ze všech knih nejblíže. K tématu je také blízko kniha *Hřbitov snů* Věroslava Mertla.

Témata, která zatím nebyla zmíněna, jsou témata 2, 4 a 6, tedy skoro třetina vytvořených tematických celků. Žádný z autorů, jejichž texty byly algoritmicky zařazeny k těmto tématům, nebyl zmíněn v žádné kapitole *Souřadnic mnohosti*. Podíváme-li se na slova tématu 2, „major“, „poručík“, „štáb“, mohli bychom usuzovat, že budou knihy zasazeny do vojenského prostředí či ho nějakým způsobem tematizovat. Když se však podíváme na jednotlivé knihy tematického shluku, uvidíme, že se tématu války či vojny týká jen kniha *Závody* Jiřího Poláka a *Jobova zvěst* Jaroslava Haidlera. V klastru jsou pak dále knihy *Milénium* Ondřeje Neffa a *Mrtvý ze Zlaté stoky* Jiřího Svejkovského. Shluk tak shromáždil spíše knihy s krimi-detektivní tématikou.

Slova tématu 4 nejsou tak lehce interpretovatelná jako slova tématu 2. Tematický shluk je proto nutné interpretovat hlavně z knih shluku. Nejblíže jsou tématu knihy *Ctitelé katastrof a Hitlerova tužka* Petra Prouzy, pak *Světlo přichází potmě* Zdeňka Rotrekla, *Opšlstisova nadace* Stanislava Komárka a *Knihovna* Jana Bažanta. Většina knih je zasazena do nedávné minulosti, nicméně další spojitosti lze nalézat jen velmi obtížně. Téma 6 je naopak vcelku čitelné. Slova jasně odkazují k románům s historickou tématikou.

8.1.3 Druhá dekáda jednadvacátého století

Důležitou poznámkou od Martina Lukáše k české próze druhé dekády jednadvacátého století bylo konstatování, že se česká literatura odklonila od tzv. „vysoké literatury“ a

přiklonila se ke stylisticky jednoduššímu psaní, lineárním dějům a výraznější

příběhovosti.[147] To je rozdíl oproti literatuře devadesátých let, kdy se těšily oblibě postmoderní postupy a vícevrstevnatá díla. Když porovnáme témata subkorpusu devadesátých let a subkorpusu dvacátých let, uvidíme, že témata jsou subkorpusu dvacátých let koherentnější a snadněji interpretovatelná. To by nasvědčovalo tomu, že jsou příběhy vystavěné na jednom tématu, který lze snáze algoritmicky zachytit. Tuto tezi však nelze nijak dokázat.

Podle Aleny Šidákové Fialové a Evy Klíčové se staly oblíbené biografické romány

zaměřující se na jednu historickou postavu.[148] Mezi takové řadí i romány, které jsou součástí korpusu, např. *Dějiny světla* Jana Němce či *Medvědí tanec* Ireny Douskové. *Dějiny světla*, román o životě fotografa Františka Drtikola, byly přiřazeny do tématu 2 k románům jako jsou *Válka zrcadel* Terezy Matouškové, *Vězněná* Pavla Renčína, ale také *Román a novely* Jana Balabána. Kniha Ireny Douskové byla naopak přiřazena do tématu 8 k románům *Svatý rváč,* biografickém románu o Jeronýmu Pražském, *Valdštejn a Lukrecie*, historickém románu o Albrechtu z Valdštejna, ale i ke knize *Nesvatý otec* Marka Dvořáka, fantasy románu zasazenému do středověku. Všechny knihy by se daly označit jako historické romány. Kdybychom se ale dívali jen na témata – „kacíř", „kněz", „boží", „církev", „arcibiskup", nabyli bychom dojmu, že se knihy týkají křesťanství a církve.

Eva Klíčová též zmiňuje obecný trend románů vypořádávajících se

s předlistopadovým režimem, které by však literární věda neoznačila za historické.[149] Žádná z v textu zmíněných knih však není v korpusu. Z témat modelu by předlistopadový režim seděl na téma 4 ke slovům jako „komunista“, „komunistický“, „stranický“ nebo „soudruh“. Kromě knih *Volavčí sítě* a *Příběhy třetího odboje* autorské dvojice Miroslava Kasáčka a Luďka Navary jsou ve shluku také dvě knihy Aleny Mornštajnové (*Slepá mapa*

a *Hotýlek*), které vypráví příběh rodinných tragédií na pozadí velkých dějin.[150] Z tohoto hlediska knihy Mornštajnové odpovídají obecnému trendu.

Literární kritik a badatel Petr. A. Bílek si všímá obecnějšího trendu české literatury,

kterým jsou krizové vztahy.[151] Ve svém textu v této souvislosti zmiňuje několik autorek – Radku Denemarkovou, Zuzanu Brabcovou, Petru Soukupovou, Annu Bolavou, Viktorii Hanišovou, Ditu Táborskou či Jakubu Katalpu. Autory zmiňuje dva, nicméně z dvojice je v korpusu pouze jedno dílo – *Stvoření* od Eugena Lišky. Díla autorek a autora jsou rozeseta napříč čtyřmi tématy, nejvíc jsou zastoupena v tématu 3, ve kterém jsou knihy *Pod sněhem* Petry Soukupové, *Malinka* Dity Táborské a *Anežka* Viktorie Hanišové. Knihy *Stvoření* Eugena Lišky a *Stopy* Zuzany Brabcové jsou v tématu 1, *Doupě* Jakuby Katalpy a *Příspěvek k dějinám radosti* Radky Denemarkové v tématu 10.

9 Diskuze

V předchozích částech jsem vysvětlila a rozebrala výsledky modelu Top2Vec. Jak se ukázalo, témata jen v ojedinělých případech korespondují s tím, jak je směřování české prózy posledních desetiletí komentováno v sekundární literatuře. V případě literatury devadesátých let dvacátého století sekundární literatura ani příliš nemluví o tématech jako takových, ale spíše o formálním (memoáry, deníky, vzpomínky), generačním (nejstarší vs. střední generace) či proudovém (střední proud) zařazení tehdejších prozaických děl. Ukázalo se, že z výsledků algoritmu Top2Vec toto rozřazení prakticky nelze vyčíst. Limitem subkorpusu devadesátých let je také fakt, že má oproti ostatním subkorpusům menší rozptyl autorů.

Výsledky zbylých dvou subkorpusů sice více korelují s tím, co tvrdí sekundární literatura, nicméně nelze říci, že by směřování literatury jednadvacátého století šlo přímo vyčíst z témat dvou modelů. Téma modelů se shoduje se sekundární literaturou tam, kde se sekundární literatura zabývá tématy ve smyslu anglického *topic* spíše než *theme*. Jasně zřetelná jsou témata týkající se relativně blízké minulosti jako téma druhé světové války

nebo komunismu. Naopak téma jako „krizové vztahy“, o kterých psali Petr A. Bílek[152] a Alena Šidáková Fialová ze slov témat tematického modelování nelze vyčíst vůbec. Model Top2Vec však častěji knihy, které v sekundární literatuře figurují jako zástupkyně téhož proudu, rovněž zařadil do stejného shluku. Stalo se tak u témat 5 první dekády, nebo tématu 10 druhé dekády jednadvacátého století. Otázkou tak zůstává, jak se témata tematického modelování, vytvořená algoritmy, vztahují ke klasické literární teorii a vědě, která vychází z vlastní terminologické a metodologické tradice.

Z posledních výzkumů vyplývá, že tematické zařazení knih k tématům modelů topic modellingu koreluje s žánrem, který jim určuje sekundární literatura, případně jiné paratexty (anotace apod.). To je závěr studie jak vědce Christofra Schöcha, který pro tematické modelování použil algoritmus LDA, tak tým vědců v čele s badatelem van

Zundertem, který použil algoritmus Top2Vec.[153] Součástí korpusu z *Českého národního korpusu* sice není informace o žánrovém zařazení knih, po bližším zkoumání titulů ve shluku se však ukázalo, že tituly přinejmenším částečně spadaly do též žánrové oblasti. Dělo se tak hlavně u historických románů a sci-fi/fantasy textů. Literatura, která se vymyká jednoznačnému žánrovému zařazení se naopak rozprostřela mezi všechna témata.

Vědecký tým van Zunderta také pozoroval, že témata často shlukují díla od stejného autora. Stejný závěr lze vyvodit i z dat v mém výzkumu. Nejvýraznější je to v subkorpusu 90. let, který obsahoval více děl od stejného autora, a kde se některé shluky vytvořily jen kolem děl jednoho autora (téma 6, 8 a 9). Jedno z vysvětlení van Zunderta a jeho kolektivu je, že se autoři či autorky často drží svého žánrového zařazení. Van Zundert

nicméně dodává, že je těžké určit, jak se k tématům tematického modelování vztahovat.[154] Díky menšímu rozsahu korpusu bylo v mém výzkumu možné zjistit, o čem knihy přibližně jsou, a tedy se více zaměřit na vztahy mezi jednotlivými tituly.

Jedním z kritérií shlukování, které vyplynulo z pozorování tematických shluků modelu Top2Vec, a které zatím nikdo v souvislosti s tematickým modelováním nezmínil, je autorský styl díla. Algoritmus Top2Vec, jak už jsem zmínila, je založen na principu přenesení textu do vektorového prostoru pomocí přístupu word2vec. Ten dokáže zachytit

jak syntaktické, tak sémantické aspekty textu.[155] Díky této vlastnosti se word2vec začal využívat ve stylometrii. Např. vědecký kolektiv kolem Benzebouchiho využil word2vec

k analýze autorství, přičemž jejich model měl úspěšnost přes 95 %.[156] Lze proto předpokládat, že i model Top2Vec zachycuje autorův či autorčin styl. Ostatně tomu nasvědčují i výsledky subkorpusu devadesátých let.

K modelům tematického modelování lze přistupovat jako k tvorbě literárního systému, který byl hojně diskutován formalisty a strukturalisty. Ti byli toho názoru, že neexistují osamocená díla, ale že díla jsou vždy zasazena do sítě literárního systému, který

tvoří soubor děl (korpus) určité epochy.[157] Především formalista Jurij Tyňanov tvrdil, že každé dílo se k systému vztahuje svým tématem, stylem a žánrem. Dílo vytržené z jednoho systému a zasazené do jiného může nabývat jiných rysů či být přiřazeno k jinému žánru.

Celý systém se pak může restrukturalizovat.[158] Takto definovaný systém by odpovídal systému vytvořenému modelem Top2Vec, který – jak jsem ukázala v této práci – shlukuje texty mimo jiné i na základě stylu a žánru, a následně počítá jejich vzdálenosti. Ty však model Top2Vec dokáže spočítat pouze uvnitř témat, nikoliv napříč tématy. Jedno téma tak tvoří malý literární systém, který může být studován.

V této části je též třeba zmínit, že jádrem této práce nebylo představit nejpřesnější model. Byla jím snaha ilustrovat vztah mezi metodou tematického modelování a klasickou literární historií / teorií. Práce má proto několik nedostatků. Především nebyly vyzkoušeny veškeré možné konfigurace parametrů algoritmů, kterými by mohlo být docíleno přesnějšího modelu.

Z některých slov v tématech (např. „malý“ ve výsledcích modelu LDA) je také patrné, že mohla být přidána na seznam stop-slov. Přílišná manipulace se stop-slovy však může zkreslit závěry podle biasu vědce či vědkyně. U některých slov (např. „jo“, „hele“ atd.) je zase sporné, zda je do seznamu zařadit, jelikož je můžeme chápat jako příznakové jazykové prostředky. Limitem práce je také omezený korpus děl, který obsahuje pouze výseč vydané knižní tvorby za dané období. Jednotlivé části textů děl jsou navíc uvnitř proházeny, což výsledky nepochybně zkresluje.

10 Závěr

V této práci jsem se zabývala tematickým modelováním současné (polistopadové) české prózy z korpusu *Českého národního korpusu,* kterou jsem rozdělila podle prvního data vydání na tři subkorpusy – díla prvně vydaná v letech 1990–1999, 2000–2009 a 2010–2018. Subkorpusy byly před modelováním očištěny od stop-slov a vlastních jmen. K modelování jsem vyzkoušela dva známé dostupné algoritmy – LDA a Top2Vec, které jsou založeny na rozdílném principu. Algoritmus LDA je založen na statistickém výskytu slov a každému dokumentu procentuálně přiřazuje témata. Algoritmus Top2Vec je založen na principu neuronových sítích a jednomu dokumentu přiřazuje pouze jedno téma. Oba algoritmy téma definují seznamem několika klíčových slov. Oběma algoritmy jsem namodelovala modely o 10, 20 a 30 tématech na dílech rozdělených na 1000 a 2000 tokenů. Model LDA jsem trénovala pouze na podstatných a přídavných jménech, model Top2Vec pak na všech slovech v očištěných subkorpusech.

Pro výběr nejpřesnějšího modelu jsem zvolila metriku skóre koherence *Cv,* která počítá, jak koherentní slova v rámci témat jsou. Modely s nejvyšším skóre koherence jsem vybrala pro následnou analýzu a interpretaci. Jelikož jsem chtěla, aby témata byla porovnatelná, skóre koherence jsem zprůměrovala napříč subkorpusy a výsledné modely jsem vybrala podle nejlepšího zprůměrovaného skóre. Pro Top2Vec byl vybraný model natrénován na dílech rozdělených na 2000tokenových částí a měl 10 témat. Vybraný LDA byl naopak natrénován na dílech rozdělených na 1000 tokenů a měl 20 témat.

Analyzovala jsem tedy jak vybraný model LDA, tak vybraný model Top2Vec. Všem modelům Top2Vec nicméně obecně metrika skóre koherence vycházela lépe, proto jsem pouze výsledky nejlepšího z nich porovnala s dosavadními poznatky literární historie, shrnuté v souborech *V souřadnicích volnosti* a *V souřadnicích mnohosti* a dalších statích.

Specifikum některých témat modelu Top2Vec bylo, že se vytvořila pouze kolem děl jednoho autora, pokud subkorpus od autora obsahoval více děl. Pokud subkorpus obsahoval alespoň dvě díla od jednoho autora, téměř vždy byla ve stejném tématu. Ve většině případů byly knihy rozděleny mezi témata vcelku rovnoměrně, což je patrné z vizualizace.

Témata modelu LDA se oproti tématům modelu Top2Vec vytvářela nerovnoměrně. U všech subkorpusů se vytvořilo jedno velké téma nebo pár větších „hlavních“ nebo „obecných“ témat, která byla přítomna z velké části u téměř všech děl. Zbylá témata se pak

vyskytovala jen buď u několika děl, nebo dokonce jen u jednoho díla. Některá slova se také často opakovala napříč tématy. Vzhledem k tomu, že je algoritmus LDA velmi závislý na vstupních datech, mohlo být opakování zapříčiněno málo, nebo naopak příliš očištěným korpusem.

Při porovnání témat se sekundární literaturou vyšlo najevo, že slova reprezentující témata odpovídala tendencím v sekundární literatuře pouze v případě, kdy téma bylo jasně definované. Tak tomu bylo v případě tématu druhé světové války nebo komunismu. Naopak témata abstraktnější jako „krize vztahů“ se ze slov témat získaných algoritmy vyčíst nedala. Korpus *Českého národního korpusu* navíc obsahuje i díla, se kterými *V souřadnicích volnosti* a *V souřadnicích mnohosti* nepracují. Zajímavé proto bylo pozorování témat, která se vytvořila pouze z knih neobsažených v souhrnech. Obecně to byly shluky kolem knih zasazených do vzdálenější historie, krimi a detektivní romány a tzv. romány pro ženy.

Některá témata sama o sobě nedávala příliš smysl. Některá sice smysl dávala, ale zpětně se ukázalo, že slova tématu jsou zavádějící. Proto pro úplné pochopení témat bylo potřeba zjistit, o čem knihy sdružené v daném tématu jsou. Pak bylo potřeba z knih zpětně reprodukovat, jakým způsobem se do témat shlukovaly. Ukázalo se, že se tematické shluky nevytvářely jen podle témat, ale i podle stylu, žánru a dalších znaků. Tematické shluky tedy nereprezentovaly pouze téma, ale kombinaci těchto literárních znaků.

Spojnice literatury též nejsou definované pouze tématem, ale i stylem, žánrem, generačním či genderovým zařazením autorstev a dalšími znaky. Jednotlivá díla tak společně tvoří síť děl, kterou – jak ukazuje tato práce – lze modelem Top2Vec vymodelovat. Závěrem mé práce s *Českým národním korpusem* tak je, že tematické modelování slouží lépe jako aproximace modelu literárního systému než jako nástroj pro hledání témat.

V kontextu těchto poznatků je důležité zdůraznit, že tematické modelování a literární teorie mohou být vzájemně obohacující. Při interpretaci výsledků tematických modelů je však vhodné zohlednit nejen samotná slova v tématech, ale i autorský styl, žánrové zařazení a další literární prvky, které hrají klíčovou roli v utváření literárních sítí. Tímto způsobem lze lépe porozumět dynamice literárního vývoje a vzájemným vztahům mezi díly a autory, což přispívá k hlubšímu chápání literárního kontextu. Bez nich by v mém výzkumu šlo jen těžko porozumět tematickým shlukům.

11 Seznam použité literatury:

Angelov, Dimo. 'Top2Vec: Distributed Representations of Topics'. arXiv, 2020. http://arxiv.org/abs/2008.09470.

Attribution 4.0 International (CC BY 4.0). 'Czech Stop Words'. Accessed 6 April 2023. https://countwordsfree.com/stopwords.

Bendík, Martin. 'Automatic Detection of Topics in Poetic Texts'. Diplomová práce, České vysoké učení technické, 2023.

Benzebouchi, Nacer Eddine, Nabiha Azizi, Nacer Eddine Hammami, Didier Schwab, Mohammed Chiheb Eddine Khelaifia, and Monther Aldwairi. 'Authors' Writing Styles Based Authorship Identification System Using the Text Representation Vector'. In *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*, 371–376. Istanbul, Turkey: IEEE, 2019.

Bílek, Petr A. 'Dekáda v české literatuře: Hodně balastu a málo velkých románů, vyniká Topol či Sidon', Accessed 30 October 2023, 27 December 2019. https://magazin.aktualne.cz/kultura/literatura/ceska-literatura-2010-2019/r~eacde10e28ca11ea8776ac1f6b220ee8/.

Blažková, Hana. 'Česká literatura nevzkvétá: Národní obrození v rukách Hosta', Accessed 27 October 2023, March 2020. https://www.advojka.cz/archiv/2020/3/ceska-literatura-nevzkveta.

Blei, David M. 'Latent Dirichlet Allocation', In *Journal of Machine Learning Research*, 2003.

Brillenburg Wurth, Kiene, and Ann Rigney, eds. *The Life of Texts: An Introduction to Literary Studies*. Amsterdam: Amsterdam University Press, 2019.

Bode, Katherine. 'Abstraction, Singularity, Textuality The Equivalence of "Close" and "Distant" Reading'. In *A World of Fiction: Digital Collections and the Future of Literary History*, 17–35. University of Michigan Press, 2019.

Cambridge University Press. *Cambridge Academic Content Dictionary*. Cambrige University Press, 2009.

Culpeper, Jonathan. 'Keywords and Characterization: An Analysis of Six Characters in Romeo and Juliet'. In *Digital Literary Studies*, edited by David L. Hoover and Kieran O'Halloran, 23–48. Routledge, 2014.

Dahllöf, Mats, and Karl Berglund. 'Faces, Fights, and Families: Topic Modeling and Gendered Themes in Two Corpora of Swedish Prose Fiction', In *DHN 2019, 4th Digital Humanities in the Nordic Countries 2019*, 2019.

Dokoupil, Blahoslav. 'Historická próza v roce nula'. In *Česká a slovenská literatura dnes*, 68–70. Praha–Opava: Ústav pro českou literaturu AV ČR, Slezská Univerzita, 1997.

Drouin, Jeffrey. 'Close- and Distant-Reading Modernism': *The Journal of Modern Periodical Studies* 5, no. 1, 110–135, 2014.

Egger, Roman, and Joanne Yu. 'A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts'. *Frontiers in Sociology* 7, 2022.

Fialová, Alena. *Česká próza po roce 1989*. Věda kolem nás. Praha: Academia, 2015.

———. 'Próza'. In *V souřadnicích mnohosti: česká literatura první dekády jednadvacátého století v souvislostech a interpretacích*, edited by Alena Fialová, Vydání první. Literární řada. Praha: Academia, 2014.

———, ed. *V souřadnicích mnohosti: česká literatura první dekády jednadvacátého století v souvislostech a interpretacích*. Vydání první. Literární řada. Praha: Academia, 2014.

Finneran, Richard J. *The Literary Text in the Digital Age*. University of Michigan Press, 1996.

Guillén, Claudio. *Mezi jedotou a růzností: Úvod do srovnávací literární vědy*. translated by Housková, Anna, Alexandra Berendová, and Mariana Housková, Praha:

Triáda, 2008.

Haman, Aleš. ‘Fikce a imaginace v prózách 90.let’. In *Česká próza 90.let 20.století*, 20–28. České Budějovice, 2002.

Hladík, Radim and et.al. *Digitální obrat v českých humanitních a sociálních vědách*, Praha: Karolinum, 2022.

Hodrová, Daniela and et. al. *... na okraji chaosu...: Poetika literárního díla 20. století*. Praha: Torst, 2001.

Hoover, David L., Jonathan Culpeper, and Kieran O’Halloran, eds. *Digital Literary Studies*. Routledge, 2014.

Iwashita, Daniela, Karel Klouda, Lucie Kořínková, and el. al. ‘Analýza motivických klastrů z oblasti aktuálních kulturně-společenských témat a jejich aplikace na materiál uměleckých textů 19. a počátku 20. století’, 2021. https://ucl.cas.cz/projekt/analyza-motivickych-klastru/.

Jockers, Matthew L., and David Mimno. ‘Significant Themes in 19th-Century Literature’. *Poetics* 41, no. 6, 2013: 750–769.

Jockers, Matthew L., and Rosamond Thalken. *Text Analysis with R: For Students of Literature*. Quantitative Methods in the Humanities and Social Sciences. Cham: Springer International Publishing, 2020.

Jockers, Matthew Lee. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. Urbana: University of Illinois Press, 2013.

Klíčová, Eva. ‘Historický román v současné české literatuře versus téma dějin v české próze (po roce 2000)’, Accessed 3 November 2023, 11 January 2019. https://www.czechlit.cz/cz/feature/historicky-roman-v-soucasne-ceske-literature-versus-tema-dejin-v-ceske-proze-po-roce-2000/.

———. ‘Kritika v osamění’, Accessed 3 November 2023, 25 January 2020. https://www.h7o.cz/clanky/12683-kritika-v-osameni.

Klimentová, Julie. ‘Texty frankofonního hip hopu z pohledu digital humanities’. Diplomová práce, Univerzita Karlova, 2022.

Křen, Michal, Václav Cvrček, Jan Hajič, Milena Hnátková, Tomáš Jelínek, and et.al. ‘SYN v9: Large Corpus of Written Czech’, 2021. http://hdl.handle.net/11234/1-4635.

Kubát, Miroslav. *Kvantitativní analýza žánrů*. Ostrava: Ostravská univerzita, Filozofická fakulta, 2016.

Machala, Lubomír. ‘Próza’. In *V souřadnicích volnosti: česká literatura devadesátých let dvacátého století v interpretacích*, edited by Petr Hruška. Praha: Academia, 2008.

Michel, JB, and et.al. ‘Quantitative Analysis of Culture Using Millions of Digitized Books’, 2011, 176–182.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. ‘Efficient Estimation of Word Representations in Vector Space’. arXiv, 2013. http://arxiv.org/abs/1301.3781.

Moretti, Franco. *Distant Reading*. London: Verso, 2013.

———. *Grafy, mapy, stromy: Abstraktní modely literární historie*. Translated by Olga Čaplyginová. Praha: Karolinum, 2014.

‘NameTag User’s Manual’. Accessed 22 September 2023. https://ufal.mff.cuni.cz/nametag/1/users-manual.

Národní knihovna ČR. ‘Personální autority’, 2022. https://aleph.nkp.cz/data/aut_ja.xml.gz.

Novakova, Iva, and Dirk Siepmann, eds. *Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives*. Springer Nature, 2019.

Panichella, Annibale. ‘A Systematic Comparison of Search-Based Approaches for LDA Hyperparameter Tuning’. *Information and Software Technology* 130, 2021. https://doi.org/10.1016/j.infsof.2020.106411.

Piorecký, Karel. 'Česká literární kultura 2001-2010'. In *V souřadnicích mnohosti: česká literatura první dekády jednadvacátého století v souvislostech a interpretacích*, edited by Alena Fialová, Vydání první. Literární řada. Praha: Academia, 2014.

Plecháč, Petr, Robert Kolár, Sylvie Cinková, Jakub Říha, and Dalibor Dobiáš. 'Korpus českého verše'. Accessed 19 August 2023. https://versologie.cz/v2/web_content/corpus.php.

Potter, Rosanne G. *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*. University of Pennsylvania Press, 1989.

Řehůřek, Radim. 'Gensim: Latent Dirichlet Allocation'. https://radimrehurek.com/gensim/models/ldamodel.html.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 'Exploring the Space of Topic Coherence Measures'. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. Shanghai China: ACM, 2015.

Schöch, Christof. 'Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama', 2021.

Sherstinova, Tatianna, and et.al. 'Topic Modelling with NMF vs. Expert Topic Annotation: The Case Study of Russian Fiction'. In *Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Mexico City, Mexico, October 12–17, 2020, Proceedings, Part II*, edited by Lourdes Martínez-Villaseñor, Oscar Herrera-Alcántara, Hiram Ponce, and Félix A. Castro-Espinoza, Vol. 12469: 134–151. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020.

Šidáková Fialová, Alena. 'Biografický román v současné české literatuře', Accessed 1 November 2023, 7 November 2018. https://www.czechlit.cz/cz/feature/biograficky-roman-v-soucasne-ceske-literature/.

———. 'Česká próza po roce 2000', Accessed 1 November 2023, 1 September 2015. https://www.czechlit.cz/cz/feature/ceska-proza-po-roce-2000/.

———. ‘Jak si stojí současné spisovatelky’. *Česká literatura*, no. 3, 2023: 364–366.

Siemens, Ray, and Susan Schreibman, eds. *A Companion to Digital Literary Studies*, Wiley-Blackwell, 2007.

Sládek, Ondřej and et. al., eds. *Slovník literárněvědného strukturalismu*. Praha, Brno: Host, 2018.

Straková, Jana, Milan Straka, and Jan Hajič. ‘A New State-of-The-Art Czech Named Entity Recognizer’. In *Text, Speech, and Dialogue*, edited by Ivan Habernal and Václav Matoušek, 8082:68–75. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

———. ‘Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition’. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 13–18. Baltimore, Maryland: Association for Computational Linguistics, 2014.

Svensson, Patrik. ‘Beyond the Big Tent’. In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 36–72. University of Minnesota Press, 2012.

Tesaříková, Anna. ‘Topic Modeling for Corpus of Czech Verse’. Bakalářská práce, České vysoké učení technické, 2022.

Tomaševskij, Boris Viktorovič. *Teorie literatury*. Translated by Renáta Štindlová and Karel Štindl. Praha: Lidové nakladatelství, 1970.

Tyňanov, Jurij. ‘Óda jako řečnický žánr’. In *Literární fakt*, 20–49. Praha: Odeon, 1988.

Uglanova, Inna, and Evelyn Gius. ‘The Order of Things. A Study on Topic Modelling of Literary Texts’, 2020.

Underwood, Ted. ‘A Dataset for Distant-Reading Literature in English, 1700-1922’, 7 August 2015. https://tedunderwood.com/2015/08/07/a-dataset-for-distant-reading-literature-in-english-1700-1922/.

Zengul, Ferhat D, Aysegul Bulut, Nurettin Oner, Abdulaziz Ahmed, Bunyamin Ozaydin, and Manju Yadav. 'A Practical and Empirical Comparison of Three Topic Modeling Methods Using a COVID-19 Corpus: LSA, LDA, and Top2Vec', *Proceedings of the 56th Hawaii International Conference on System Sciences*, 2023.

Zizler, Jiří. 'Otevřená dekáda'. In *V souřadnicích volnosti: česká literatura devadesátých let dvacátého století v interpretacích*, edited by Petr Hruška. Praha: Academia, 2008.

Změlík, Richard. *Konceptualizace barev v narativní fikci na pozadí kvantitativních modelů*. Olomouc: Univerzita Palackého v Olomouci, 2019.

Zundert, Joris J van, Marijn Koolen, Julia Neugarten, Peter Boot, Willem van Hage, and Ole Mussmann. 'What Do We Talk About When We Talk About Topic?', *Proceedings http://ceur-ws. org ISSN 1613*, 2022.

12 Seznam obrázků:

13 Seznam tabulek:

[100] Tatianna Sherstinova and et.al., 'Topic Modelling with NMF vs. Expert Topic Annotation: The Case Study of Russian Fiction', in *Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Mexico City, Mexico, October 12–17, 2020, Proceedings, Part II*, ed. Lourdes Martínez-Villaseñor et al., vol. 12469, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2020).

[101] Joris J van Zundert et al., 'What Do We Talk About When We Talk About Topic?', 2022.

[102] Christof Schöch, 'Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama', 2021.

[103] Tesaříková, 'Topic Modeling for Corpus of Czech Verse'; Bendík, 'Automatic Detection of Topics in Poetic Texts'.

[104] Bendík, 'Automatic Detection of Topics in Poetic Texts', 36–37.

[105] Bendík, 38.

[106] Bendík, 62.

[107] Bendík, 62.

[108] Tesaříková, 'Topic Modeling for Corpus of Czech Verse', 29–30.

[109] Bendík, 'Automatic Detection of Topics in Poetic Texts', 62.

[110] Julie Klimentová, 'Texty frankofonního hip hopu z pohledu digital humanities' (diplomová práce, Praha, Univerzita Karlova, 2022).

[111] Klimentová, 87.

[112] *Digitální Obrat v Českých Humanitních a Sociálních Vědách*, 2022.

[113] Hladík, 'Modelování témat v české sociologii: typy autorství a citační ohlas v odborných textech'.

[114] Hladík, 166.

[115] Hladík, 173.

[116] Hladík, 176–177.

[117] Sherstinova and et.al., 'Topic Modelling with NMF vs. Expert Topic Annotation: The Case Study of Russian Fiction'; Bendík, 'Automatic Detection of Topics in Poetic Texts'.

[118] Jockers and Mimno, 'Significant Themes in 19th-Century Literature'; Dahllöf and Berglund, 'Faces, Fights, and Families: Topic Modeling and Gendered Themes in Two Corpora of Swedish Prose Fiction'; Hladík, 'Modelování témat v české sociologii: typy autorství a citační ohlas v odborných textech'.

[119] Inna Uglanova and Evelyn Gius, 'The Order of Things. A Study on Topic Modelling of Literary Texts', 2020; Tesaříková, 'Topic Modeling for Corpus of Czech Verse'; Bendík, 'Automatic Detection of Topics in

Poetic Texts'.

[120] Schöch, 'Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama'; van Zundert et al., 'What Do We Talk About When We Talk About Topic?'

[121] Attribution 4.0 International (CC BY 4.0), 'Czech Stop Words', accessed 6 April 2023, https://countwordsfree.com/stopwords.

[122] Národní knihovna ČR, 'Personální autority', 2022, https://aleph.nkp.cz/data/aut_ja.xml.gz.

[123] Jana Straková, Milan Straka, and Jan Hajič, 'A New State-of-The-Art Czech Named Entity Recognizer', in *Text, Speech, and Dialogue*, ed. Ivan Habernal and Václav Matoušek, vol. 8082, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer Berlin Heidelberg, 2013), 68–75; Jana Straková, Milan Straka, and Jan Hajič, 'Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition', in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland: Association for Computational Linguistics, 2014), 13–18.

[124] 'NameTag User's Manual', Accessed 22 September 2023, https://ufal.mff.cuni.cz/nametag/1/users-manual.

[125] Šidáková Fialová, 'Jak si stojí současné spisovatelky', 365.

[126] Radim Řehůřek, 'Gensim: Latent Dirichlet Allocation', https://radimrehurek.com/gensim/models/ldamodel.html.

[127] V této implementaci se parametr „beta" nazývá „eta"

[128] Röder, Both, and Hinneburg, 'Exploring the Space of Topic Coherence Measures'.

[129] Tesaříková, 'Topic Modeling for Corpus of Czech Verse'; Bendík, 'Automatic Detection of Topics in Poetic Texts'.

[130] Tesaříková, 'Topic Modeling for Corpus of Czech Verse', 29.

[131] Bendík, 'Automatic Detection of Topics in Poetic Texts', 61.

[132] Uglanova and Gius, 'The Order of Things. A Study on Topic Modelling of Literary Texts', 66.

[133] Obě témata obsahují jméno Fišerová, z čehož lze mimo jiné usuzovat, že nástroj NameTag neodchytil veškerá jména v korpusu. Vzhledem k tomu, že se jména v tématech téměř neobjevují, je tato nedokonalost přehlédnutelná.

[134] Nikde není bohužel dohledatelné, podle jakého klíče jsou knihy do korpusu vybírány. Na svých wiki stránkách však *Český národní korpus* zmiňuje, že se vědci snažili o žánrovou vybalancovanost.

[135] Machala, ‘Próza’, 281.

[136] Machala, 289.

[137] Machala, 286.

[138] Machala, 288–291.

[139] Machala, 293.

[140] Machala, 296.

[141] Machala, 297.

[142] Alena Fialová, ed., *V souřadnicích mnohosti: česká literatura první dekády jednadvacátého století v souvislostech a interpretacích*, Vydání první, Literární řada (Praha: Academia, 2014), 361.

[143] Fialová, ‘Próza’, 344.

[144] Fialová, *V souřadnicích mnohosti*, 354.

[145] Fialová, 347.

[146] Fialová, ‘Próza’, 350.

[147] Klíčová, ‘Kritika v osamění’.

[148] Šidáková Fialová, ‘Biografický román v současné české literatuře’.

[149] Klíčová, ‘Historický román v současné české literatuře versus téma dějin v české próze (po roce 2000)’.

[150] Hana Blažková, ‘Česká literatura nevzkvétá: Národní obrození v rukách Hosta’, Accessed 27 October 2023, March 2020, https://www.advojka.cz/archiv/2020/3/ceska-literatura-nevzkveta.

[151] Bílek, ‘Dekáda v české literatuře: Hodně balastu a málo velkých románů, vyniká Topol či Sidon’.

[152] Bílek.

[153] Schöch, ‘Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama’; van Zundert et al., ‘What Do We Talk About When We Talk About Topic?’

[154] van Zundert et al., ‘What Do We Talk About When We Talk About Topic?’, 406.

[155] Tomas Mikolov et al., ‘Efficient Estimation of Word Representations in Vector Space’ (arXiv, 6 September 2013), http://arxiv.org/abs/1301.3781.

[156] Nacer Eddine Benzebouchi et al., ‘Authors’ Writing Styles Based Authorship Identification System Using the Text Representation Vector’, in *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)* (2019 16th International Multi-Conference on Systems, Signals & Devices (SSD), Istanbul, Turkey: IEEE, 2019), 371–376.

[157] Ondřej Sládek and et. al., eds., ‘Systém’, in *Slovník literárněvědného strukturalismu* (Praha, Brno: Host, 2018), 730–734.

[158] Jurij Tyňanov, ‘Óda jako řečnický žánr’, in *Literární fakt* (Praha: Odeon, 1988), 20–49.

Příloha 1. Top2Vec 1990–1999

Normalizace, Vztahy	**Mafie**	**Krajina**	**Viewegh**	**Vaculík**
socialismus	jo	uvidet	usmev	decko
dokazat	hele	zvedat	naprosto	chlapec
pritel	tenhle	okno	roman	rukopis
predevsim	prachy	ucitit	pohlednout	slunko
zaroven	chlapek	okamzik	pochopitelne	takovyto
urcity	kseft	otcuv	hlavolam	morce
nazor	holka	uslyset	obvykle	pravit
usili	jenze	snih	zajezd	psat
otazka	povidat	tma	pripadat	vanoce
pripadat	koukat	svetlo	trochu	pomery
trebaze	kamos	kdesi	tvarit	podivit
predstava	ksicht	dotek	puzzle	doma
presvedceni	takovyhle	komandant	pozorovat	vcera
zcela	von	priblizit	fotrik	minit
budoucnost	chlap	brno	zasmat	dovedet
duvod	kouknout	hladina	segra	cist
mozny	nejaky	krok	samozrejme	domu
naprosto	auto	stin	dokonce	dopis
alespon	policajt	sklo	vazne	zapis
nyni	akorat	usta	doslova	takto

Vlastimil Třešňák	**Historické romány**	**Bohumil Hrabal**	**Ladislav Klíma**	**Autencitní literatura**
times	turnovsky	cikanka	vzkriknout	celba
the	turnov	kracet	zvolat	teho
inu	grunt	lis	jmout	kurva
sajdkara	hrabe	balik	on	sichta
kriknout	vrchnost	vrchni	sestricka	mas
ehm	rozeny	sklep	bandita	tym
kostkovany	nybrz	ohromny	uzrit	predak
vytisteny	rod	dvur	buh	pica
ruksak	obec	dzban	jako	srat
hospodar	osoba	nazpatek	vskutku	synek
plastikovy	uvadet	divat	strasny	rikat
street	predek	koryto	velet	bracha
plnovous	chram	snoubenec	uchopit	ju
zakricet	hejtman	pruvan	ucinit	chlapik
kozeny	soused	host	ulehnout	kluk
ukazovak	ponevadz	optat	zajiste	dul
tuzka	panstvi	malicky	cetnik	sachta
zastera	kopa	cikan	stanout	hospoda

plateny	vdat	hotel	milostivy	tabak
televizor	syn	papir	zarvat	pivo

Příloha 2. Top2Vec 2000–2009

Rodinné vztahy	**Detektivní/ Krimi**	**Dětský příběh**	**Dějiny**	**Každodenní život**
matcin	ponekud	trava	zcela	jit
maminka	informace	koukat	rochlice	pekny
laska	major	okolo	tehdy	zase
okno	kontakt	spacak	knihovna	jet
obraz	porucik	dlan	znamy	vzit
prestat	stab	chechtat	jaksi	domu
milovat	minuta	kluk	mesto	scenar
ulice	mauss	batoh	dejiny	chodit
kvet	vyuzit	povidat	nyni	spat
doopravdy	jednat	vsude	kniha	jist
maly	delo	jo	velmi	jezdit
lhostejny	akce	strcit	ovsem	delat
otcuv	zalezitost	vzduch	politicky	pusa
cizi	teren	plny	clen	zavolat
ted	cinnost	sirem	pocatek	tata
kdysi	pocitac	hele	trestanec	reziser
strach	kapral	bricho	namesti	vydrzet
poslouchat	kosmicky	zase	stoleti	kobyla
dovest	policie	reka	nejen	doma
zase	dispozice	zada	rada	rici

Historické romány	**Román pro ženy, oddychové čtení**	**Dospívání, vzpomínky na totalitní režim**	**2. světová válka**	**Zvíře vs. člověk**
markyz	polibit	fakt	zidovsky	vlci
markyza	mobil	jo	arabsky	vlk
panos	pohlednout	hele	izraelsky	smecka
madame	mama	furt	jeruzalem	cosi
chlum	usmat	ponevadz	palestinec	dotknout
purkrabi	prekvapene	blby	ghetto	vzhuru
prokurator	taska	tata	fiserova	beztak
nadvori	reditelka	basnicka	palestinsky	prestoze
sluha	rozesmat	celkem	zabijet	sotva
pisar	obejmout	skola	rabin	znovu
holstejn	zadivat	dost	transport	sam
vlkodlak	usmev	delat	reznik	srst
kralovsky	doktor	takovyhle	valka	pach

olomoucky	nahle	tenhle	zid	otcuv
kupec	vazne	nejaky	britsky	vystrel
kocar	prominout	prdel	terorista	zahledet
biskup	prikyvnout	uplne	lagr	balvan
zdobeny	vypnout	proboha	vychod	hrbitov
saty	kabelka	tridni	kasarna	stacit
opacit	zeptat	zkratka	nemecko	pohreb

Příloha 3. Top2Vec 2010–2018

Tělesnost	**Historické romány**	**Rodinné vztahy**	**Komunismus**	**Prázdninové**
cítit	tma	jo	komunista	tata
pohnout	citadela	mama	maminka	strejda
pomalu	soumrak	gauc	komunisticky	ves
ucítit	temnota	prachy	stranicky	chalupa
upir	nebe	holka	soudruh	letos
dlan	stín	fakt	vojna	naves
prst	bledy	uplne	odbojar	prazdniny
pritisknout	cosi	tata	socialisticky	taska
zaseptat	tvor	hele	ceskoslovensky	lhotka
ret	svetlo	koupelna	rezim	autobus
krk	padat	telefon	mamincin	chata
dotknout	kamen	sednout	socialismus	cerveny
zadivat	hluboky	jit	republika	bracha
oprit	krev	hrozne	soudruzka	deda
ucho	bezejmenny	chvilka	skola	mama
prudce	kost	doma	rodic	bouda
vterina	demon	mobil	ceskoslovensko	pamatovat
tise	uprostred	zenska	hotylek	silnice
dech	dole	brecet	tatinek	kolo
uvedomit	cítit	bavit	mazak	strejdanek

Mafie	**Psaní**	**Církev**	**Fantasy/ Středověk**	**????**
lobbista	spisovatel	kacir	rytir	bohyne
cching	roman	knez	panacek	kopanice
sef	autor	bozi	podmesti	vlastovka
investor	literarni	cirkev	mag	telo
veta	napriklad	arcibiskup	brneni	policistuv
konwicky	ctenar	mistr	prilba	ptacek
solarni	sexualni	svaty	kulka	postak
tema	text	krcma	ocelovy	vyzkum
zdroj	sex	uceni	zbran	suong

moskva	napsat	farar	magie	doupe
ministr	ich	biskup	citadela	joga
menit	basnik	modlit	mestsky	nekolikrat
padelek	vypravec	copak	morgain	hradiste
projekt	autoruv	milost	zbroj	jednotlivy
firma	kritik	kazat	budova	ritual
byznys	povidka	koncil	mrtvy	dlan
obraz	psat	hrich	strilet	nemoc
katedra	manzelka	baryton	carodejka	peclive
text	zrejme	septat	krev	cinnost
konference	popisovat	kostel	vystrelit	proces

Příloha 5.LDA 1990-1999

Téma 1	**Téma 2**	**Téma 3**	**Téma 4**	**Téma 5**
okno	doba	malý	manželka	město
malý	duše	okno	ředitel	syn
tělo	přítel	bůh	rodič	dcera
noc	chlapec	zeď	práce	doba
stůl	malý	město	škola	babička
město	krásný	pravý	doba	smrt
krásný	pravda	sen	pravda	jméno
láska	vlastní	modrý	malý	rod
ulice	konec	kapsa	žák	známý
Doba	pocit	rameno	učitel	ulice

Téma 6	**Téma 7**	**Téma 9**	**Téma 8**	**Téma 10**
doktor	peníze	práce	poručík	sovětský
maminka	doktor	město	město	agent
pivo	stůl	Moskva	právo	tajný
stůl	práce	dopis	císař	služba
kluk	doba	Alexander	noc	zpráva
kurva	telefon	Alexandrovič	vlastní	americký
krásný	fotřík	vysoký	malý	stát
práce	auto	malý	chvilka	doba
papír	holka	vedoucí	syn	jméno
Praha	pravda	doba	pravda	vyzvědač

Téma 11	**Téma 12**	**Téma 13**	**Téma 14**	**Téma 15**
kniha	hlavolam	český	přítel	šaman
komandant	puzzle	velvyslanec	srdce	stan
papír	práce	grunžijský	římský	tábor
balík	část	Praha	bůh	lovec
lis	úsměv	rukopis	láska	náčelník
sklep	řešení	ministr	doba	duch
plný	kniha	doba	město	veliký
práce	máma	úřad	verš	zvíře
knížka	otázka	ministerstvo	kniha	les
cikánka	vražda	pravý	krásný	jeskyně

Téma 16	**Téma 17**	**Téma 18**	**Téma 19**	**Téma 20**
morče	babička	specialista	autobus	komandant
inženýr	dědeček	autobus	žebřík	český
stůl	učitelka	bůh	noc	jméno
banka	divadlo	hospodář	ředitel	známý
kolega	škola	Ježek	pouť	šavle
chlapec	slečna	banda	malý	předek
malý	teta	podlaha	městečko	stůl
klec	malý	tma	vlastní	hrabě
kočka	zákop	dlaň	jaffský	ústa
zvíře	Ničín	halda	doba	malý

Příloha 6. LDA 2000-2009

Téma 1	**Téma 2**	**Téma 3**	**Téma 4**	**Téma 5**
malý	malý	smrt	máma	kluk
doba	okno	malý	doba	doba
okno	stůl	bůh	práce	soudruh
noc	doba	tělo	noc	bratr
stůl	tělo	doba	malý	máma
pravda	práce	pocit	láska	ředitel
ulice	láska	jméno	město	škola
plný	kuchyně	rodina	případ	holka
Praha	noc	válka	doktor	sovětský
jméno	pravda	noc	peníze	stůl

Téma 6	**Téma 7**	**Téma 8**	**Téma 9**	**Téma 10**
malý	kůň	Chlum	ocas	dlaň
auto	pes	velitel	holka	táta
práce	auto	tatínek	doba	stůl
noc	les	královský	malý	obličej
holka	kolo	panoš	slečna	rameno
město	práce	kluk	pravda	záda
ženská	doba	sestra	Praha	prst
pes	malý	malý	práce	tělo
peníze	vůz	kapitán	profesorka	auto
krásný	město	domov	máma	okno

Téma 11	**Téma 12**	**Téma 13**	**Téma 14**	**Téma 15**
dělo	Fišerová	družice	mrtvý	dozorce
plukovník	kluk	počítač	major	smrt
kapitán	kufr	kosmický	kapitán	studna
major	holka	firma	pacient	kniha
Sermonte	noc	Brian	stůl	Luxemburgová
štáb	židovský	kapitán	poručík	šachový

italský	německý	pes	stoka	kosmoděmjanský
kóta	bratr	skafandr	vůz	konec
děda	válka	skrín	jméno	studně
Talián	láska	dráha	fotografie	Rosa

Téma 16	**Téma 17**	**Téma 18**	**Téma 19**	**Téma 20**
doba	moře	pes	vzácný	Bezděk
pravda	řeka	kolo	milostný	síran
kamarád	malý	auto	planeta	malý
známý	pivo	přívěs	záchodový	doba
zub	chlap	krásný	malý	noc
řeka	škola	rodina	dojem	stůl
malý	sochař	týden	doba	okno
kůň	doba	zbytek	plný	láska
jistý	fotr	právo	výsledek	smrt
kilometr	pravý	práce	víra	ulice

Příloha 7. LDA 2010-2018

Téma 1	**Téma 2**	**Téma 3**	**Téma 4**	**Téma 5**
tělo	malý	malý	malý	táta
doba	Praha	tělo	tělo	máma
kniha	škola	stůl	okno	malý
vlastní	doba	okno	město	auto
román	práce	rameno	ulice	holka
kůň	válka	světlo	noc	stůl
láska	kluk	víla	prst	práce
jméno	stůl	doba	pocit	rodič
práce	okno	krev	doba	doba
příběh	pravda	tma	auto	okno

Téma 6	**Téma 7**	**Téma 9**	**Téma 8**	**Téma 10**
doba	bratr	základna	malý	rada
malý	loď	doba	práce	koule
sluneční	malý	skupina	syn	ohnivý
šéf	král	Brno	auto	doba
rodina	tělo	archiv	doba	soudruh
ulice	opat	tajný	vlastní	stůl
případ	doba	vězení	hotýlek	loď
stůl	sestra	zbraň	peníze	informace
slunce	smrt	člen	týden	vlastní
práce	Eadulf	bezpečnostní	stůl	jistý

Téma 11	**Téma 12**	**Téma 13**	**Téma 14**	**Téma 15**
táta	kůň	malý	sklo	chalupa

máma	malý	vlk	auto	stůl
doba	páter	kapitán	máma	ves
zeď	kolo	divoch	malý	okno
mrtvý	noc	vlastní	stůl	auto
okno	Vsetín	manžel	dědek	kuchyně
plný	tělo	kůň	okno	strejda
bůh	hradní	stan	sklář	táta
malý	bůh	ostatní	kára	pole
noc	plný	bůh	káva	světlo

Téma 16	**Téma 17**	**Téma 18**	**Téma 19**	**Téma 20**
student	plukovník	voják	řeka	holka
doba	ministr	rota	Kanaďan	řeka
stůl	Praha	mazák	Ptáčník	film
práce	auto	velitel	homunkulus	režie
kniha	vnitro	soudruh	doba	Paříž
katedra	otázka	četa	třída	Kanaďan
text	lobbista	vojna	holka	doba
profesor	stůl	Janovice	sloup	Fidel
ulice	malý	náčelník	lavice	Viewegh
historik	práce	kapitán	výtah	dík

Příloha 4. Top2Vec skóre koherence

subkorpus	**Počet tokenů**	**Počet témat**	**Skóre koherence**
1990_1999	**2000**	**10**	**0,519283119**
1990_1999	1000	20	0,502778138
1990_1999	2000	30	0,496573158
1990_1999	2000	20	0,493719789
1990_1999	2000	37	0,490496554
1990_1999	1000	59	0,484620111
1990_1999	1000	30	0,484470576
1990_1999	1000	10	0,47905946

subkorpus	**Počet tokenů**	**Počet témat**	**Skóre koherence**
2000_2009	2000	48	0,472325054
2000_2009	1000	30	0,468513239
2000_2009	1000	73	0,467452038
2000_2009	2000	20	0,464438974
2000_2009	2000	30	0,464314972
2000_2009	**2000**	**10**	**0,458386459**
2000_2009	1000	20	0,443196361

2000_2009	1000	10	0,443087131

subkorpus	Počet tokenů	Počet témat	Skóre koherence
2010_2019	2000	52	0,479528656
2010_2019	**2000**	**10**	**0,460548681**
2010_2019	1000	30	0,45982375
2010_2019	2000	20	0,452957332
2010_2019	1000	20	0,447683958
2010_2019	2000	30	0,443875699
2010_2019	1000	86	0,441232939
2010_2019	1000	10	0,436673738

Příloha 4. LDA skóre koherence

subkorpus	Počet tokenů	Počet témat	Skóre koherence
1990_1999	**1000**	**20**	**0,428313615**
1990_1999	2000	30	0,413391561
1990_1999	2000	20	0,412138611
1990_1999	1000	30	0,41128449
1990_1999	2000	10	0,358091371
1990_1999	1000	10	0,347738659

subkorpus	Počet tokenů	Počet témat	Skóre koherence
2000_2009	2000	20	0,389793663
2000_2009	**1000**	**20**	**0,388840204**
2000_2009	1000	30	0,366856752
2000_2009	2000	30	0,35301361
2000_2009	2000	10	0,33379914
2000_2009	1000	10	0,292472216

Subkorpus	Počet tokenů	Počet témat	Skóre koherence
2010_2019	1000	30	0,411237024
2010_2019	2000	30	0,40807349
2010_2019	2000	20	0,386911791
2010_2019	**1000**	**20**	**0,386498548**
2010_2019	1000	10	0,337314891
2010_2019	2000	10	0,298921085